Teaching Case

Teaching Data Literacy Using Titanic Survival Factors

Mark Sena
Xavier University
sena@xavier.edu

Thilini Ariyachandra
Xavier University
ariyachandrat@@xavier.edu

Xavier University
Cincinnati, OH 45207

Abstract

The Titanic disaster is a topic that continues to fascinate. As the importance of analytics continues to grow in industry, data literacy skills have become increasingly important in business education. This project allows students to use the passenger data from the Titanic to build their data literacy skills using an engaging, experiential topic. The project requires students to extract, transform, describe, analyze, and draw conclusions regarding the factors that impacted survival on the Titanic. The project can be deployed using various application software and tools. We describe how the assignment can be completed using Excel, Tableau, and Python using the Pandas library.

Keywords: experiential learning, analytics, data literacy, Tableau, Python, Pandas, Excel

1. INTRODUCTION

The world today is characterized by a data revolution, with an unparalleled volume of data being generated, disseminated, and scrutinized every day. Data has emerged as a vital asset in the digital age, influencing individuals, businesses, and society as a whole. It offers valuable insights that can steer decision-making, enhance performance, and foster innovation (Mandinach & Gummer, 2016). For individuals in society, data is instrumental in improving personal decision-making and lifestyle management. As per a Pew Research Center survey, 69% of U.S. adults monitor at least one health, fitness, or diet metric, underscoring the significance of data in personal health management (Fox & Duggan, 2013). Additionally, a study by the National Bureau of Economic Research revealed that personal finance apps, which utilize data to offer financial insights, can help decrease spending by 15.7% (Baker, Meyer, Pagel, & Yannelis, 2018).

In the business sphere, data is crucial for strategic planning, operational efficiency, and gaining a competitive edge. Companies that base their decisions on data and analytics, rather than intuition or gut feeling, perform better across almost all metrics. A study conducted by Harvard Business School and MIT Sloan School of Management on over 2,000 public companies across various industries found that data-driven companies had superior financial performance, were more likely to survive, and were more innovative (Garg, 2022). A recent study by

McKinsey discovered that data-driven companies outperform their competitors by up to 20% (Garg 2022). At the societal level, data can inform policy-making, enhance public services, and drive social change. The United Nations has highlighted the importance of data for achieving the Sustainable Development Goals, with a call for a "data revolution" to tackle global challenges such as poverty, inequality, and climate change (United Nations, 2014). Furthermore, a Pew Research Center study found that 65% of Americans believe that government use of data to make decisions can improve public services (Rainie & Anderson, 2017).

Given its impact on the individual, business and society, the ability to understand and utilize data effectively has become a highly sought after skill in industry. As society has grown more dependent on data, it has become imperative to ensure that all individuals possess the necessary skills to be data literate. Adopting data literacy throughout the organization has begun to gain attention as a means of ramping up data driven decision making in day to day as well as strategic operations in companies. All data related roles are growing at a rapid pace with data scientists roles expected to grow at 36 percent from 2021 to 2031 according to the U.S. Bureau of Labor Statistics (2023). According to recent study of expert interviews and survey responses from over 1,200 global C-level executives and 6,000 employees conducted by Qlik (2023), data literacy, will be the most sought-after skill by 2030. Furthermore, 85 percent of executives predict that data literacy will be as crucial in the future as computer literacy is in the present.

As the demand for data literacy skills continue to grow it falls upon institutions of higher education to arm future employees with the necessary data savvy required in organizations. This data savvy, also known as data literacy, empowers individuals to collect, manage, evaluate, and apply data competently to decision making (Mandinach & Gummer, 2016). Teaching data literacy skills has its own challenges. Many students express concerns and face challenges in acquiring data literacy skills. One of the primary issues students face is a lack of confidence in their abilities to interpret and analyze data. According to a study by Prinsloo and Slade (2017), students often feel overwhelmed by the complexity of data and the technical skills required to manipulate and interpret it. This lack of confidence can lead to a fear of data and a reluctance to engage with it. Another common issue is the lack of understanding of the relevance of data literacy to their future careers. Many students fail to see how data literacy skills will be applicable in their chosen fields, leading to a lack of motivation to acquire these skills (Koltay, 2015). The project described in this paper attempts to introduce data literacy skills in an interesting and entertaining way that can ease students into the world of data manipulation.

2. OVERVIEW OF DATA LITERACY

In recent years, various definitions of data literacy have emerged. It is loosely considered the ability to understand and use data (Frank, Walker, Attard & Tygel, 2016). According to UNESCO (2004) "literacy" is the ability to define, understand, interpret, create, and calculate through relevant written and printed sources. More specifically, data literacy can be described as the ability to read, work with, analyze, and argue with data (Koltay, 2015). Some include the need for statistical analysis as part of data literacy. Dag (2019) describes it as knowing how to access data in different ways, asking questions, and making basic statistical analysis (Dağ, 2019). Past literature has also described it in term of actions as well as in terms of specific competencies that an individual should acquire. Pothier and Condon (2019) identified seven baseline data literacy competencies to help prepare business students for the workforce: (1) data organization and storage, (2) understanding data used in business contexts, (3) evaluating the quality of data sources, (4) interpreting data, (5) data-driven decision making, (6) communicating and presenting effectively with data, and (7) data ethics and security.

Upon investigating multiple definitions of data literacy, Wolff, Gooch, Cavero Montaner, Rashid and Kortuem (2016) developed a comprehensive definition for data literacy as "the ability to ask and answer real-world questions from large and small data sets through an inquiry process, with consideration of ethical use of data. It is based on core practical and creative skills, with the ability to extend knowledge of specialist data handling skills according to goals. These include the abilities to select, clean, analyze, visualize, critique and interpret data, as well as to communicate stories from data and to use data as part of a design process" (pg 23). The definition describes seven abilities that involve data literacy:(1) selecting, (2) cleaning, (3) analyzing, (4)visualizing, (5)critiquing, (6)interpreting data, as well as (7) communicate stories. The following definition by D'Ignazio and Bhargava (2015) express data literacy in terms of the ability to carry out four specific actions related to data: reading data, working with data,

analyzing data, and arguing with data. It is widely used in widely as a concise and easy definition that can be incorporated in to business practices. The following project enables students to gain experience in nearly all of the preceding foundational abilities.

3. PROJECT DESCRIPTION

Objectives and Questions for the Analysis

To analyze the Titanic data set, we are interested in learning what factors were associated with survival. These include such questions as:
1) Were females more likely than males to survive?
2) Were children or senior citizens more likely to survive than adults?
3) Were first or second class passengers more likely to survive than third class passengers?
4) Were passengers whose cabin was located in certain locations more likely to survive?
5) Were passengers who paid a higher fare more likely to survive than those who paid a lower fare?
6) In the movie Titanic, the main character Rose survived without a lifeboat by clinging on to a board. Could she have been based on a real passenger?

Extracting the Data

The Titanic data is widely available on the internet and has been used for analytical projects for many years. Kaggle.com (kaggle, 2023) is one common source. Note that there are various subsets of the data for projects that may differ in terms of number of records and in file type. The data set that we use in our project includes 1309 passengers. The dataset includes fourteen fields as depicted in Table 1.

For a dataset that is over one hundred years old, it is surprisingly rich. However, there are various missing data points throughout the data. The age and cabin locations of passengers is more complete for first and second class passengers than for third class passengers.

Transforming the Data Set

To conduct the analyses in Table 1, we need to first create some new columns that will allow us to create charts and statistics. These include:
1) creating a numeric equivalent of the gender field.
2) creating a label equivalent of the survived field (so that charts can show "survived" or "died" instead of 0, 1 as labels)
3) creating a variable that groups age into Blank (equal to ""), children (under 18), adults, senior citizens (over 64)
4) creating a one letter (LEFT() function) cabin location variable

Field	Description
Survival (1, 0)	1 = Survived, 0 = Died
PClass (1,2,or 3)	Passenger Class (first, second or third class)
Name	Name of Passenger
Sex (male, female)	Gender of Passenger
Age (numeric or blank if unknown)	Age of Passenger (if known)
Sisp	Number of Siblings/Spouses aboard
Parch	Number of Parents/Children aboard
Ticket	Ticket Number
Cabin (blank if unknown)	Cabin Number
Embarked (C, Q, or S)	Port of Embarkment: Cherbourg, Queenstown, Southampton
Boat (blank if passenger not on lifeboat)	Lifeboat number or letter
Body (blank if not recovered)	Body Number if recovered
Fare (in Pounds)	Fare Paid
Home/Dest	Home town and destination of passenger

Table 1: Fields in Titanic Data Set

To make these transformations in Excel, we would use simple IF functions for the first two steps, a nested IF() function for the third, and a LEFT() function for step four. In Tableau, these transformations could be done in Excel or using various actions in Tableau itself (such as creating a calculated field variables or grouped field). In Python, after reading the dataset, students can create list variables by looping through each data then creating new fields by setting the data frame field equal to the list variable.

Demographics and Summary Data

Before exploring the relationships, students should perform summarize key passenger data to establish a base line. These include:
1) Showing the number of survivors vs deaths
2) Showing the number of each gender on board
3) Showing the number of children, adults, and senior citizens on board
4) Showing the number of passengers in each class (1,2,3)
5) Showing the number of passengers in various fare groups.

6) Showing the number of passengers in each cabin class

In Excel, students could create pivot tables for each column shown above and use Count as the Value Field Setting. See Figure 1 in the Appendix for an example of the output in Excel. In Tableau, a similar approach could be used with worksheets showing the Count measure in a row or column and a field (such as Survived) in the row or column to produce a column or bar chart. All six summary charts could also be combined into a dashboard visualization. In Python, there are various summary code options in the Pandas library such as histogram plots (df.hist(‘variable)) or tabular results using code such as .counts or .size (df.groupby(‘variable’).size()

Relationship with Survival

Now that students have a baseline of passenger characteristics, we can examine how each factor relates to survival. These include:
1) Showing the average survival rate by gender
2) Showing the average survival rate by age group
3) Showing the average survival rate by passenger class
4) Showing the average survival rate broken down by fare groups.
5) Showing the average survival rate by cabin level.

In Excel, students can create pivot tables to show tabular and visualizations by selecting the survived column with Average as the Value Field setting with other fields in the Row. See Figure 2 in the Appendix for an example. In Tableau, students would simply drag the Survived field in the row or column and the other field in the opposite row or column. Students would need to adjust the Measure option for Survived from the Sum default to Average. In Python, students could use the groupby option in the Pandas library for each combination. For example, the following code would show survival by gender: print(df.groupby('gender')['survived'].mean())

Exploring Multiple Survival Factors

Since there are several possible combinations of factors that could be chosen, it would be preferable to develop an interactive visual to allow the user to explore relationships. For example a user may wish to view survival rate of adult females in first class. To do so in Excel, students can create a pivot table with various slicers that can be selected to interactively filter the data. We ask students to show both the survival rate (average of Survived) and the number of passengers (count of Survived). See Figure 3 in the Appendix for an example. In Tableau, similar approach could be used with Filters. In Python, it is more difficult in the Pandas library to create interactive statistics without teaching students to develop a user interface. However, students can explore several combinations of options using the groupby and mean. For example the following would show survival rate by gender and passenger class: print(df.groupby(['gender','pclass'])['survived']. mean()).

Correlation Between Survived and Factors

To better understand the survival factors, students can compute the correlation between survival and each numeric field in the data set to measure the strength and direction of the linear relationships. In Excel, this can be conducted using the CORREL function (and other possible methods). Students can also create a chart comparing the relationships. See Figure 4 in the Appendix for an example of correlations that use conditionally formatted bar charts. In Tableau, students can create Scatter Plots with a trend line added that shows both the correlation and p-value of the relationships. In python, a simple df.corr() line of code will produce a table of correlations among all numeric fields in the data set.

Exploring Whether “Rose” Could Be Real

To engage the students, especially those who have seen the movie, we ask students to examine potential passenger characteristics that meet those of the character, Rose, who was a young adult female who survived without making it onto lifeboat. In Excel, students can convert the primary data into a table then filter the fields to meet the criteria. This would include filters such as Female, Age between 15 and 35, survived = 1 and boat equal to blank (or null). See Figure 4 in the Appendix for an example. In tableau, the students can drag the Name field (and any other field of interest) into the Row, then uncheck Aggregate Measures, and add each relevant field into the Filter box and select appropriate ranges or allowable values. In Python, this would be most easily done using a for loop with an if statement that includes all combinations of appropriate criteria and an indented statement to print name and other fields.

Explaining Results and Findings

To summarize and interpret the results of the analysis, we ask students to include a written narrative addressing the six questions previously identified as objectives and research questions. These could be embedded in worksheets in Excel, as separate Word documents or as Comments in Python. See Figure 5 in the Appendix for an example of the Findings

4. CONCLUSIONS AND LIMITATIONS

Limitations of this assignment include the potential for academic dishonesty (including the possible use of AI tools) and the popularity of the data set in higher education. Because students are analyzing the same data with little possibility in variation of results, the potential for collaboration across students is high. AI resources such as ChatGPT (and many other emerging AI) are becoming a challenge and opportunity for educators. ChatGPT is quite adept at producing the results in this assignment or in giving step by step instructions or code to students, which can take much of the critical thinking out of the assignment. Because the data set is so easily accessed, students may encounter similar assignments in other courses.

Extensions of the assignment could include more advanced analytics and visualizations with statistical tests, multivariate analyses, or storyboarding and/or video presentations. There are also additional fields that were not used in the analysis and other datasets that could be used in tandem with the Titanic passenger data to extend the scope of the project. A more advanced approach could include the analysis of incomplete data by imputation. For the Python version of the analysis, AI tools could be used to generate the code. The dataset could also be used for skills testing rather than as an assignment. A survey of student satisfaction could also provide valuable feedback on the analysis.

In summary, the Titanic passenger data set is a simple yet interesting resource that students can easily analyze in varying levels of complexity. We present an assignment that is engaging and appropriate for introductory or intermediate data literacy for students in business analytics, statistics or introductory programming courses. The assignment can be deployed with various software applications, including our examples of Excel, Tableau, and Python. Other common examples could include PowerBI, R, SAS, or other programming languages or statistical software.

5. REFERENCES

Baker, S., Meyer, S., Pagel, M., & Yannelis, C. (2018). How Does Household Spending Respond to an Epidemic? Consumption During the 2020 COVID-19 Pandemic. National Bureau of Economic Research.

Dağ, H. (2019). Data literacy competencies and training. Journal of Librarianship and Information Science, 51(3), 692-702.

D'Ignazio, C., & Bhargava, R. (2015). Approaches to building big data literacy. In Proceedings of the Bloomberg Data for Good Exchange Conference.

Fox, S., & Duggan, M. (2013). Tracking for health. Pew Research Center's Internet & American Life Project.

Frank, M. R., Walker, J., Attard, J., & Tygel, A. F. (2016). Data literacy: What it is and how to promote it. In Proceedings of the 3rd Open Data Research Symposium.

Garg, R. (2022). The impact of data-driven decision-making on firm performance. McKinsey Quarterly.

Kagle, (2023). The Complete Titanic Dataset. https://www.kaggle.com/datasets/vinicius150987/titanic3. Accessed September 15,2023

Koltay, T. (2015). Data literacy: in search of a name and identity. Journal of Documentation, 71(2), 401-415.

Mandinach, E. B., & Gummer, E. S. (2016). Data literacy for educators: Making it count in teacher preparation and practice. Teachers College Press.

Pothier, S., & Condon, M. (2019). Data literacy for business students: A case study in a Canadian university. Journal of Business & Finance Librarianship, 24(3-4), 147-166.

Prinsloo, P., & Slade, S. (2017). Big data, higher education and learning analytics: Beyond justice, towards an ethics of care. In Big Data and Learning Analytics in Higher Education (pp. 275-292). Springer.

Qlik. (2023). Data Literacy: The Upskilling Evolution. Qlik and The Future Labs.

Rainie, L., & Anderson, J. (2017). The future of truth and misinformation online. Pew Research Center.

UNESCO (2004) United Nations Educational, Scientific, and Cultural Organization. The plurality of literacy and its implications for

ISSN: 2473-4901
v9 n5979

policies and programs. Paris, France: UNESCO Press.

United Nations. (2014). A world that counts: Mobilising the data revolution for sustainable development. UN Data Revolution Group.

U.S. Bureau of Labor Statistics. (2023). Occupational Outlook Handbook. U.S. Bureau of Labor Statistics.

Wolff, A., Gooch, D., Cavero Montaner, J. J., Rashid, U., & Kortuem, G. (2016). Creating an understanding of data literacy for a data-driven society. The Journal of Community Informatics, 12(3).

ISSN: 2473-4901
v9 n5979

6. APPENDICIES

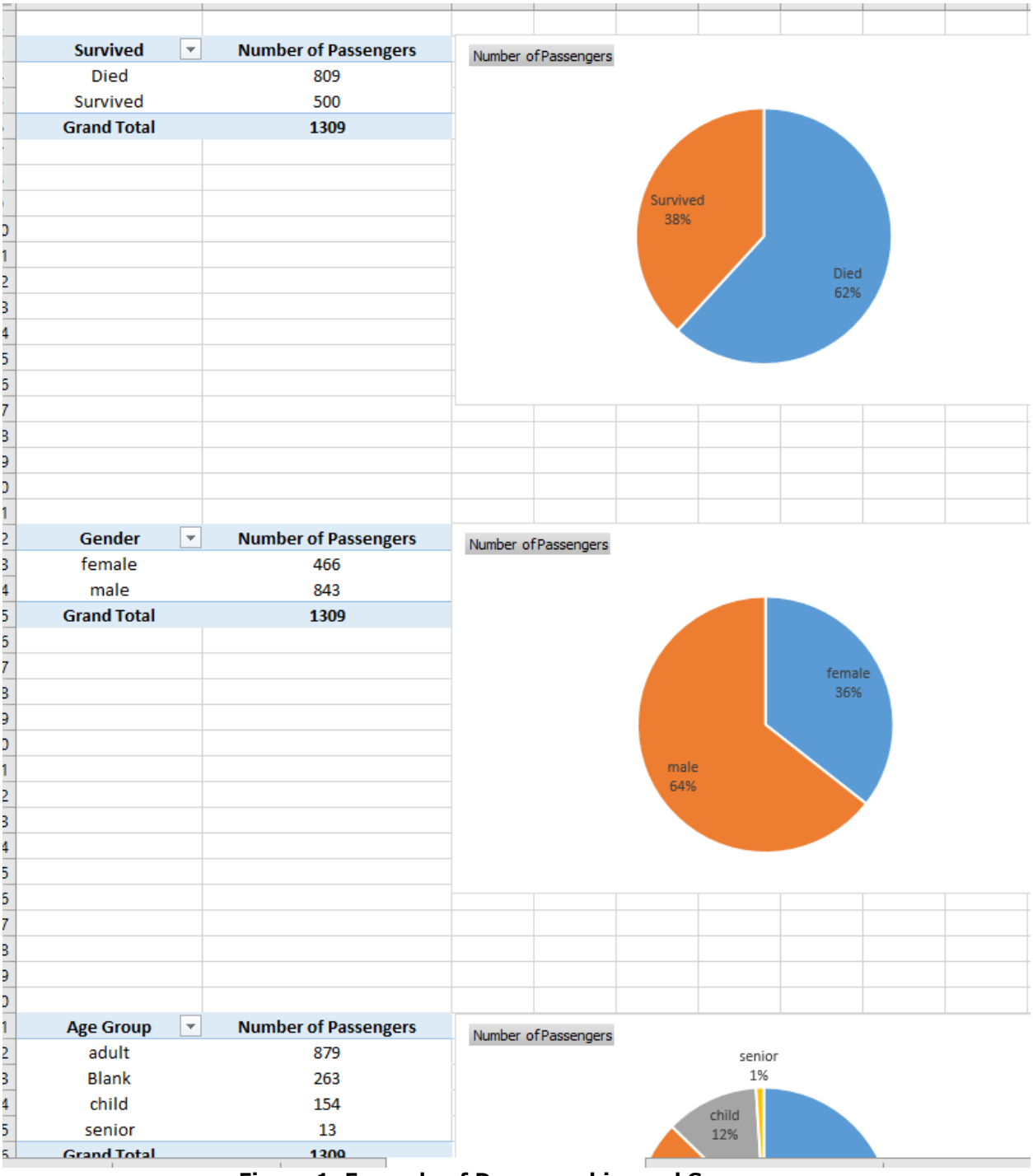


Figure 1: Example of Demographics and Summary

ISSN: 2473-4901
v9 n5979

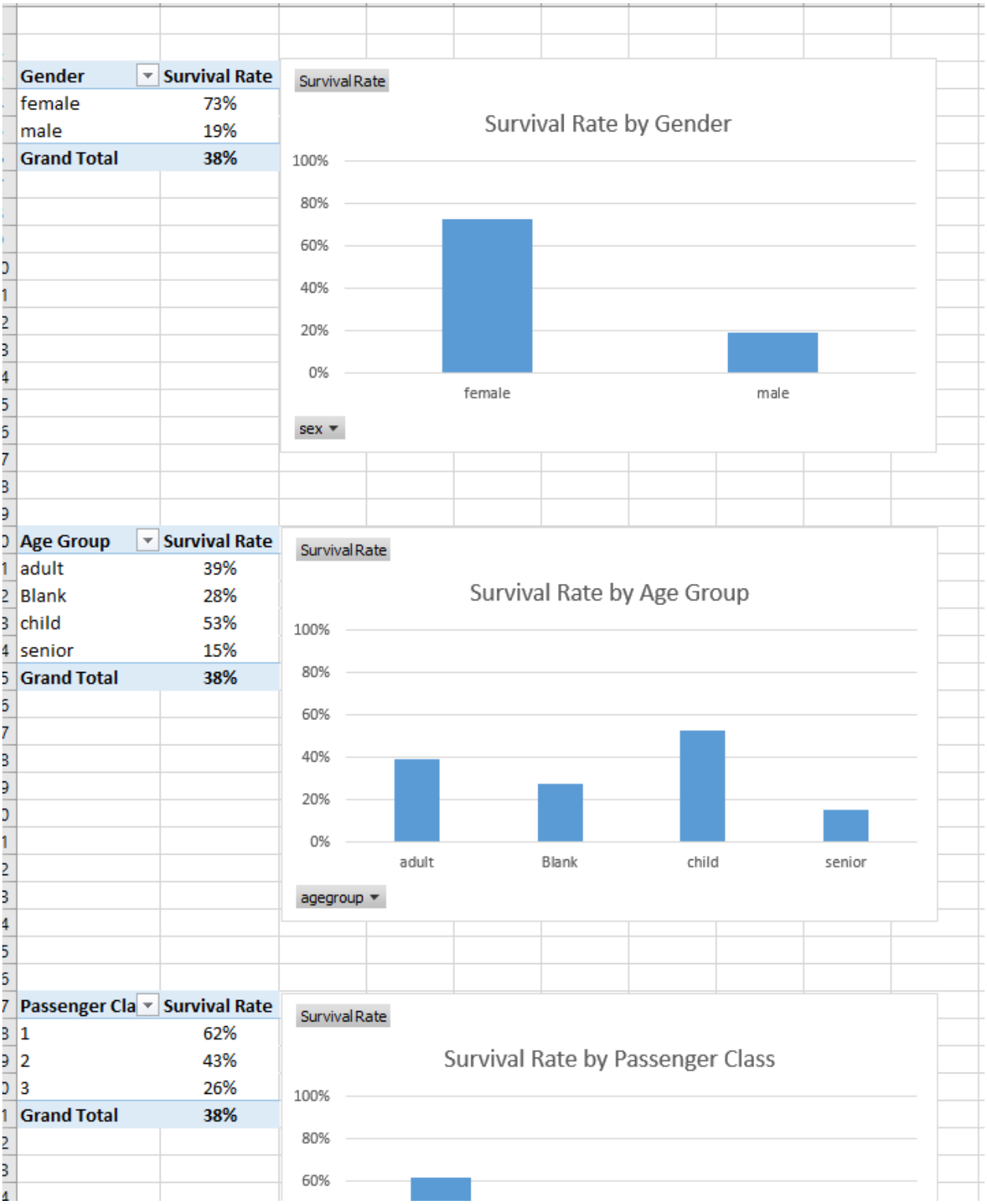


Figure 2: Example of Relationships with Survival

ISSN: 2473-4901
v9 n5979

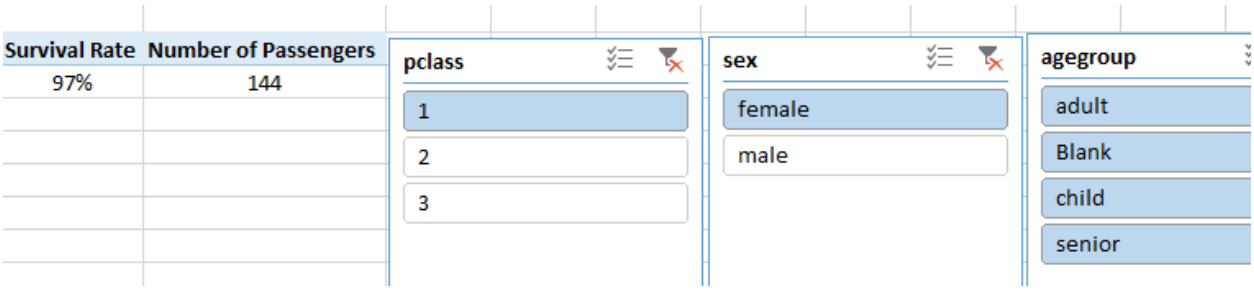
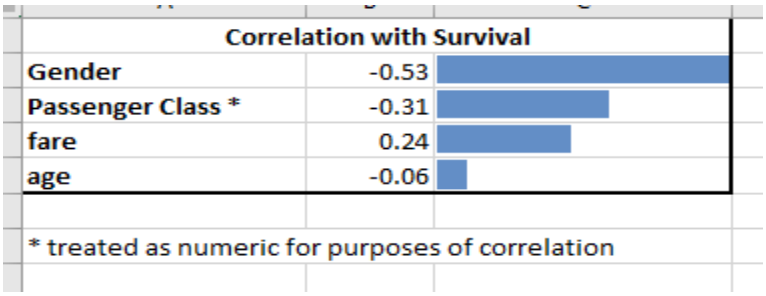


Figure 4: Example of Exploring Survival Factors



Correlation with Survival	
Gender	-0.53
Passenger Class *	-0.31
fare	0.24
age	-0.06

* treated as numeric for purposes of correlation

Figure 3: Example of Correlation Between Survival and Other Fields

pclass	survived-alpha	name	sex	age	agegroup	boat	home.dest
2	Survived	Doling, Miss. Elsie	female	18	adult		Southampton
2	Survived	Doling, Mrs. John T (Ada Julia Bone)	female	34	adult		Southampton
2	Survived	Ilett, Miss. Bertha	female	17	child		Guernsey
2	Survived	Nasser, Mrs. Nicholas (Adele Achem)	female	14	child		New York, NY
2	Survived	Renouf, Mrs. Peter Henry (Lillian Jefferys)	female	30	adult		Elizabeth, NJ
2	Survived	Trout, Mrs. William H (Jessie L)	female	28	adult		Columbus, OH
3	Survived	Backstrom, Mrs. Karl Alfred (Maria Mathilda Gustafsson)	female	33	adult		Ruotsinphytaa, Finland New York, N
3	Survived	Drapkin, Miss. Jennie	female	23	adult		London New York, NY
3	Survived	Heikkinen, Miss. Laina	female	26	adult		
3	Survived	Honkanen, Miss. Eliina	female	27	adult		
3	Survived	McGowan, Miss. Anna "Annie"	female	15	child		
3	Survived	Osman, Mrs. Mara	female	31	adult		
3	Survived	Yasbeck, Mrs. Antoni (Selini Alexander)	female	15	child		

Figure 4: Example Table of Possible "Rose" Passengers

	A
1	**Findings**
2	38 percent of the 1309 passengers survived the Titanic
3	The passengers were predominatly Male (64%) and Adults (67%) or age unkown (20%) Veryy few passengers were children (12%) or seniors (1%)
4	53% of the passengers were listed as 3rd class compared to 21% 2nd class and 25% first class
5	Most passengers paid 0 to 50 for their fare and most did not have a cabin listed
6	
7	Females survived at much higher rate than Males (73% to 19%)
8	Children survived at a higher rate (51%) than other age groups
9	First class passengers survived at higher rate (62%) than 2nd class (43%) and 3rd class (26%) passengers
10	Similarly, passengers who paid the lowest fare of less than 50 survived at a lower rate (32%) than those who paid a higher fare
11	
12	In exploring multiple factors, we were able to identify, for example, that the 144 first class, female passengers survived at a 97% rate
13	By contrast, the 664 males in 2nd or 3rd class survived at only a 15% rate
14	
15	The correlation statistics show that gender is most strongly related to survival, followed by passenger class and fare. Age was not strongly related to survival.
16	
17	After filtering for passengers who were female, survived, and had a lifeboat that was blank, there were 13 records that met the criteria
18	In looking at the remaining passengers, there were two passengers who were in 2nd class and were not married
19	These include Miss. Elsie Doling and Miss. Bertha Ilett
20	

Figure 5: Example of Summary and Findings

ISSN: 2473-4901
v9 n5979

To analyze the Titanic data set, we are interested in learning what factors were associated with survival. These include such questions as:

1) Were females more likely than males to survive?
2) Were children or senior citizens more likely to survive than adults?
3) Were first or second class passengers more likely to survive than third class passengers?
4) Were passengers whose cabin was located in certain locations more likely to survive?
5) Were passengers who paid a higher fare more likely to survive than those who paid a lower fare?
6) In the movie Titanic, the main character Rose survived without a lifeboat by clinging on to a board. Could she have been based on a real passenger?

To conduct the analyses above, we need to first create some new columns using the IF function that will allow us to create charts and statistics. These include:

1) creating a numeric equivalent of the gender field.
2) creating a label equivalent of the survived field (so that charts can show "survived" or "died" instead of 0, 1 as labels)
3) creating dummy variables for children (under 18) and senior citizens (over 64)
4) creating a one letter (left function) cabin location variable

Before exploring the relationships, we should perform some demographic pivot tables that show counts of to describe the passengers. These include:

1) Showing the number of survivors vs deaths
2) Showing the number of each gender on board
3) Showing the number of children, adults, and senior citizens on board
4) Showing the number of passengers in each class (1,2,3)
5) Showing the number of passengers in various fare groups.
6) Showing the number of passengers in each cabin class

Now that we have a baseline of these demographics, we can examine how each factor relates to survival. We can do this by developing pivot tables and charts that include:

1) Showing the average survival rate by gender
2) Showing the average survival rate by age group
3) Showing the average survival rate by passenger class
4) Showing the average survival rate broken down by fare groups.
5) Showing the average survival rate by cabin level.

To explore multiple factors above, we can create a pivot table and chart that includes:

1) Showing the average survival rate and count of survivors with slicers to allow interactive filtering for each of the other factors (gender, age, class, fare, cabin level)

To better understand the relationships among the data, we can use the CORREL function to:

1) Create a chart showing the correlation between survival rate and each of the numeric or dummy variables

ISSN: 2473-4901
v9 n5979

To explore that final question, we can convert our data sheet into an interactive table to allow:
1) Filtering the list of passengers to include the characteristics of Rose (female, age range, survived, no lifeboat)

Lastly, we can write a short summary answering the six questions at the top of this worksheet.

In order to complete the analysis, you should have a worksheet for each of the colored sections above, including:
1) the raw data (Titanic) with new columns and later modified to include an interactive table with filters (yellow sections)
2) the demographics of the passengers (green)
3) relationships with survival (orange)
4) interactive exploration of survival factors (blue)
5) chart of correlations (gray)
6) a short summary of findings (brown)

Figure 6: Student Instructions (for Excel)