I pledge my honor that I have abided by the Stevens Honors System. ***Krisin Kim***

1. Indicate data quality issues in the dataset? Give specific examples of the issue from the data.

The given record data has several quality issues of following:

a. **Incompleteness,** since several cells have a blank entry, unknown data
b. **Incorrectly matched attributes**/ values for “first” attributes. Attribute’s name infers the values to be the first name of people, but given values are categorized better with the “full name with prefix” attribute.
c. The values for the “first” name attribute are **not in one type** because some values involve “(“ and “)” symbols.
d. **Accuracy issues** in values of “first” attribute since looking at some values like Mr. Anders Johan, for “last name” of Andersson, I assume this to be incorrectly typed or collected.
e. **Inconsistency** in data representation of numbers. The decimal places of values of the “fare” attribute are inconsistent, varying from 0 decimal to 4 decimals
f. Addressing the values like “Mr. Anders Johan”, each attribute representing a simple value (ex. Last name) rather than **composite value** (ex. name) will enhance the data quality.
g. Additionally, regarding the believability of the data, there should be certain assumptions made before analyzing the data since the data itself doesn’t validate the credibility of it

2. Indicate data type of each of the attributes (e.g., Nominal, Ordinal, Symmetric/Asymmetric Binary, Ratio, Interval).

1. **Last: Nominal**
    a. Names of people
2. **First: Nominal**
    a. **Names of people**
3. **Gender: Symmetric Binary**
    a. The outcomes are equally important
4. **Age: Ratio**
    a. Inherent zero point, can speak of values as being an order of magnitude
5. **Class: Ordinal**
    a. Classes are picked from 1, 2, or 3, values representing first class, second class, or third class.
6. **Fare: Ratio**
    a. Inherent zero point, can speak of values as being an order of magnitude
7. **Embarked: Nominal**
    a. Names of cities
8. **Survived: Asymmetric Binary**
    a. The outcomes are not equally important since we focus on data of people who survived

3. What characteristics are shared by all passengers whose fare is 0? Should they be considered when analyzing fare statistics?

- They share two attributes: gender (M) and embarked (Southampton)
- If the values were nulls, I would consider ignoring the values, however, I consider them meaningful in analyzing the data since the value is 0 which representing $0, not the unknown.

4. Which embarkation city had the lowest-paying passengers on average?

Southampton had the lowest paying passengers on average.

- Based on the result from Google Sheet query,
  `=QUERY(F:G,"select G, avg(F) group by G")`

embarked	avg fare
Cherbourg	59.95414405
Queenstown	13.27602987
Southampton	27.24365139

5. How many married women over age 50 embarked in Cherbourg? (Married women are denoted by "Mrs.")

4 people

- Obtained the result from
  `=COUNTIFS(C:C, "=F",E:E, ">50", G:G, "=Cherbourg", B:B, "*Mrs.*")`

6. What is the most common last name among passengers? What is the average number of passengers per last name?

9 people with last name of "Andersson"
Average: about 1 person (1.335832084)

Obtained the result from
`=QUERY(Titanic!A2:A892,"select A, count(A) group by A")`

	A	B
1		count
2	Abbing	1
3	Abbott	2
4	Abelson	2
5	Adahl	1
6	Adams	1
7	Ahlin	1
8	Aks	1
9	Albimona	1
10	Alexander	1
11	Alhomaki	1
12	Ali	2
13	Allen	2
14	Allison	3
15	Allum	1
16	Andersen-Jense	1
17	Anderson	1
18	Andersson	9
19	Andreasson	1
20	Andrew	1
21	Andrews	2
22	Angle	1
23	Appleton	1
24	Arnold-Franchi	2

(cropped only the partial to display)

Then, used
`=AVERAGE(B2:B)`

7. What's the survival rate for passengers in the three different classes, i.e., what fraction of passengers in each class survived? Find the answer using spreadsheet functions only - don't perform arithmetic by hand!

1st class: 63%
2nd class: 47%
3rd class: 25%

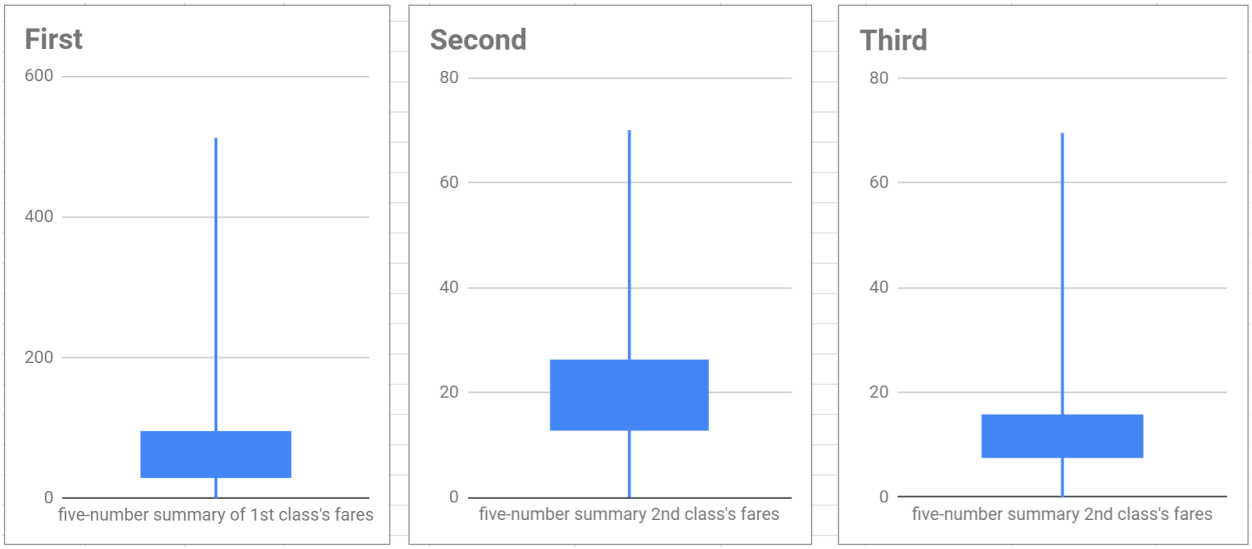
Obtained the result from

```
=QUERY(A:J,"select D, count(D) group by D")
=QUERY(A:J,"select D, count(D) where H='yes' group by D")
```

class	count class		# survived	survival rate
	0	class	count class	
1	216	1	136	0.6296296296
2	184	2	87	0.472826087
3	491	3	119	0.2423625255

8. Create a box plot showing the five number summary of fares paid by passengers in each class. The three bars should be labeled "*first*", "*second*", "*third*". You may follow steps in Excel Box Plot Chart (Links to an external site.) to build a box-plot.

Three box plots display the five number summary, which is a set of descriptive statistics that provides information about a dataset, of fares paid by passengers in each class, 1st, 2nd and 3rd class. End points of vertical line are a minimum and a maximum of the data set, the ends points of the box in the middle capture the 1st and 3rd quartile, and the median locates in the blue box.



Followed the steps below to obtain the plots above

1. Ordered the data by class (1,2,3)
2. Calculated the five-number summary for fares of each class by functions in Google Sheets, min(), max() and quartile()

class	min	Q1	median	Q3	max
1	0	30.92395	60.2875	93.5	512.3292
2	0	13	14.25	26	70
3	0	7.75	8.05	15.5	69.55

3. Then, created a box plot for each class using candlestick chart type

9. Create a stacked bar chart showing the number of passengers in each class, divided into male and female (i.e., three bars).

Each stack is composed of number of female and male in a different color, and each stack represent a gender population of each class. The ratio of female and male in class 1 and class 2 seem even with bit less female in each class, however in class 3, the number of male take more than double of the number of female.

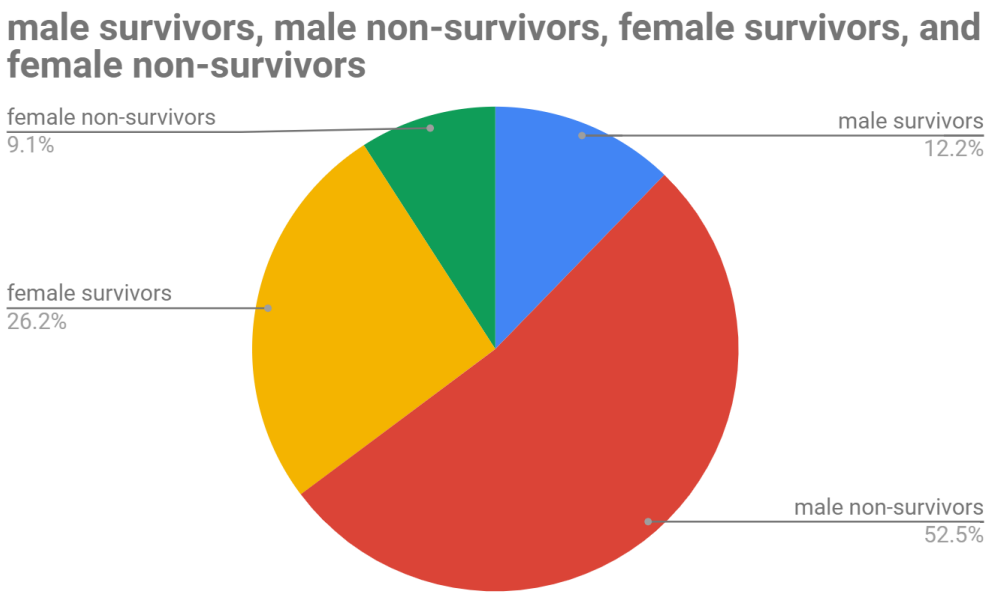
Obatined the result from

```
=QUERY(A:K,"select D, count(D) where C='F' group by D")
```

	Male		Female
class	count class	class	count class
1	122	1	94
2	108	2	76
3	347	3	144

Then, created a stacked bard chart with the result above from query

10. Create a pie chart showing the relative number of male survivors, male non-survivors, female survivors, and female non-survivors (i.e., four slices).



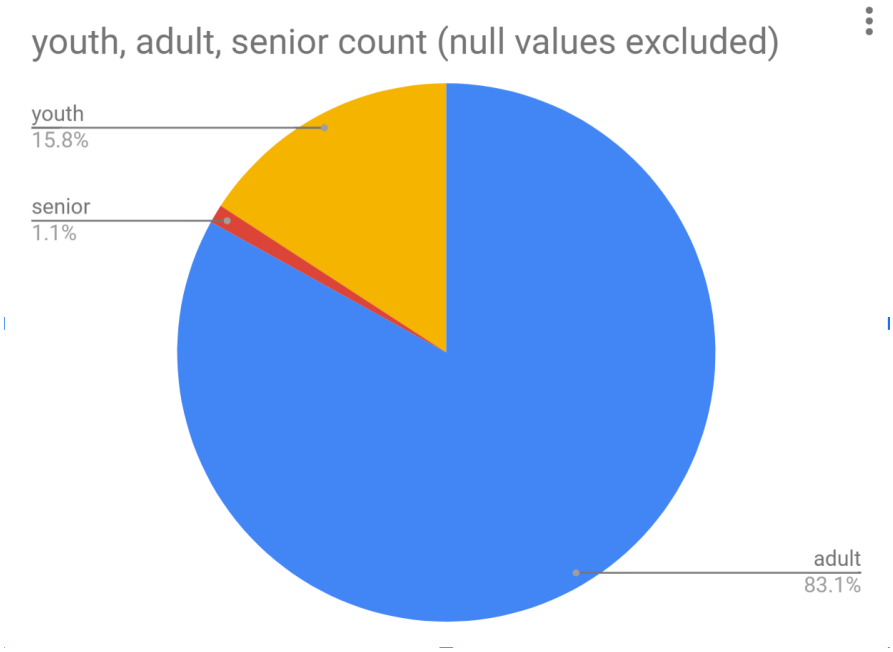
Obtained the result from,

- male survivors: `=countifs(C:C,"M", H:H,"yes")`
- male non-survivors: `=countifs(C:C,"M", H:H,"no")`
- female survivors: `=countifs(C:C,"F", H:H,"yes")`
- female non-survivors: `=countifs(C:C,"F", H:H,"no")`

male survivors	109
male non-survivors	468
female survivors	233
female non-survivors	81

11. [Bonus] Let "youth" denote passengers whose age is under 18, "adult" denote passengers age 18-59, and "senior" denote passengers whose age is 60 and above. Create a pie chart with four slices showing the relative number of youth, adult, senior, and those whose age is unknown. Hint: consider using function =countifs() in Excel

The pie chart shows that the majority of passengers fall under adult category, being between 18 and 59 years old (inclusive). However, this doesn't include passengers with their age unknown.



Obtained the result from,

```
=ifs(AND(E2>0,E2<18) ,"youth",and(18<=E2,E2<=65), "adult",E2>65, "senior")
```

```
=QUERY(I2:I, "select I, count(I) group by I")
```

#N/A	177
adult	593
senior	8
youth	113

Data:
https://docs.google.com/spreadsheets/d/1aNGEcji2gh2N9JGEWFv0p9XNwoGL80V6ChYVOsCdfDw/edit#gid=1608974636