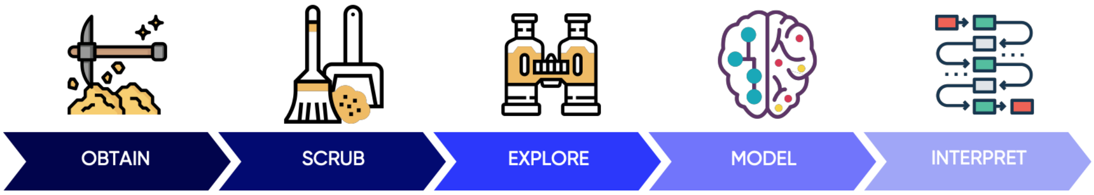
Minilab 4b Worksheet

Categorical Data Analysis

We have seen that the Data Science Lifecycle involves the following phases.



In this minilab, we will look at exploratory data analysis with categorical variables, working our way through obtaining/loading the data, scrubbing/cleaning the data into a tidy form, and exploring the data.

Step 1. Obtain the Data

The *RMS Titanic* was a passenger ship that famously sank in the North Atlantic in 1912 after hitting an iceberg.

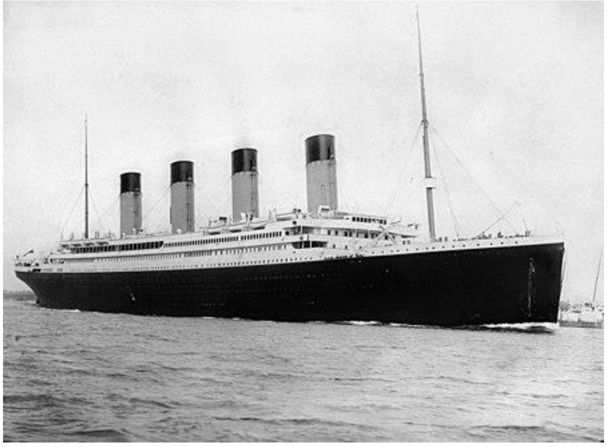


Image from Wikipedia

(1) The data we will use comes from the R package titanic and contains 891 rows (passengers) and 12 columns (variables). This is a subset of passenger data from the actual Titanic voyage. The R code below loads the data and converts it into a tibble (tidy table).

```
library(tidyverse)
#install.packages("titanic")    # run only once
library(titanic)
titanic = as_tibble(titanic_train)
titanic
```

The variables given are:

PassengerID	
Survived	0=no, 1=yes
Pclass	1=first, 2=second, 3=third
Name	
Sex	female/male
Age	
SibSp	number of siblings/spouses aboard
Parch	number of parents/children aboard
Ticket	ticket number
Fare	passenger fare paid (British pounds)
Cabin	
Embarked	port ship boarded (C=Cherbourg, Q=Queenstown, S=Southampton)

Notice that what we are most interested in is which passengers survived (the variable *Survived*) but notice that this is a categorical variable.

Step 2. Scrub

(2) Our analysis will only focus on some of the variables. We will also rename the Sex variable as Gender.

```
titanic = select(titanic,Survived,Pclass,Name,Gender=Sex,Age,Fare)
titanic
summary(titanic)
```

Notice that summary() gives min/LQ/median/mean/UQ/max for Survived and Pclass.

Warning! – This indicates that R is treating those two variables as *quantitative* variables.

(3) Survived (0=no, 1=yes) and Pclass (1=first, 2=second, 3=third) make more sense as *categorical* variables. We convert the *type* of the variable from quantitative (literally they are numbers on loading the data) to categorical (called factors in R).

```
titanic$Survived = as_factor(titanic$Survived)
titanic$Pclass = as_factor(titanic$Pclass)
titanic
summary(titanic)
```

Now we see from summary() that Survived and Pclass are summarised by **counts** of how many passengers rather than mean, median, etc.

This is how we can easily see which variables are quantitative and which are categorical.

(4) Notice that there are some NA (“not available”) values in the dataset indicating *missing values*. For example, we see it in the Age variable. This is the best way to record when a value is missing, rather than using a specific value like 9999.

Step 3. Explore

(5) To explore a dataset, a good starting point is a *scatter matrix* which shows relationships between the quantitative variables.

```
library(GGally)
ggscatmat(titanic)
```

In this case we only see Age and Fare (with a weak correlation) since these are the only two quantitative variables.

(6) We want to look at how the categorical are variables related. Clearly Name is unique to each passenger, so we don't much want to do any data analysis on it, but we might want to know who are the passengers that meet some conditions (data queries using filter and select as in previous minilabs).

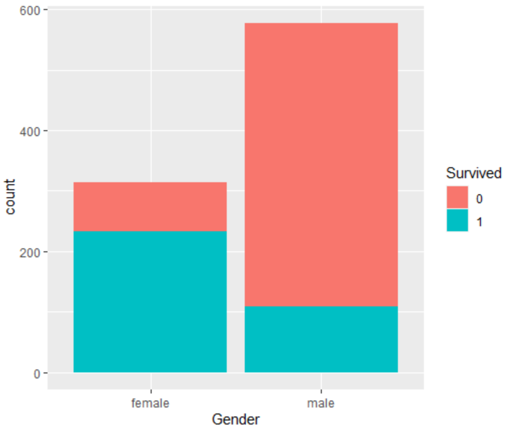
Remember that the dataset has data on 891 passengers.

```
# Survived
summary(titanic$Survived)
```

The output shows that 549 died and 342 lived.

(7) Now we break down survival by Gender. Note that the colours are chosen by R.

```
# Survived by Gender
ggplot(titanic, aes(x=Gender)) +
  geom_bar(aes(fill=Survived))
```



What do you notice from this graphical plot?

Write a dplyr pipe to extract these counts, using group_by() and summarise().

(8) We can also break down survival by passenger class (Pclass). Note that 1=first class (wealthiest passengers), 2=second class, 3=third class (poorest passengers). It is easy to copy-and-paste these plots from RStudio to Word using the Export menu (just above the plot next to Zoom) and Copy to Clipboard, then finally paste into Word.

```
# Survived by Pclass
ggplot(titanic, aes(x=Pclass)) +
  geom_bar(aes(fill=Survived))
```

What do you notice from this graphical plot?

Write a dplyr pipe to extract these counts, using group_by() and summarise().

(9) The previous two plots have shown how survival is related to gender and how survival is related to passenger class (treating gender and passenger class independently). Now we can break down survival by both gender and passenger class, i.e., women in first class, men in first class, women in second class, men in second class, women in third class, and men in third class.

```
# Survived by Gender and Pclass
ggplot(titanic, aes(x=Gender)) +
  geom_bar(aes(fill=Survived)) +
  facet_wrap(~Pclass)
```

Based on this graphical plot, describe or summarise the relationship between Survived, Gender and Class. Which bars in the plot are most striking?

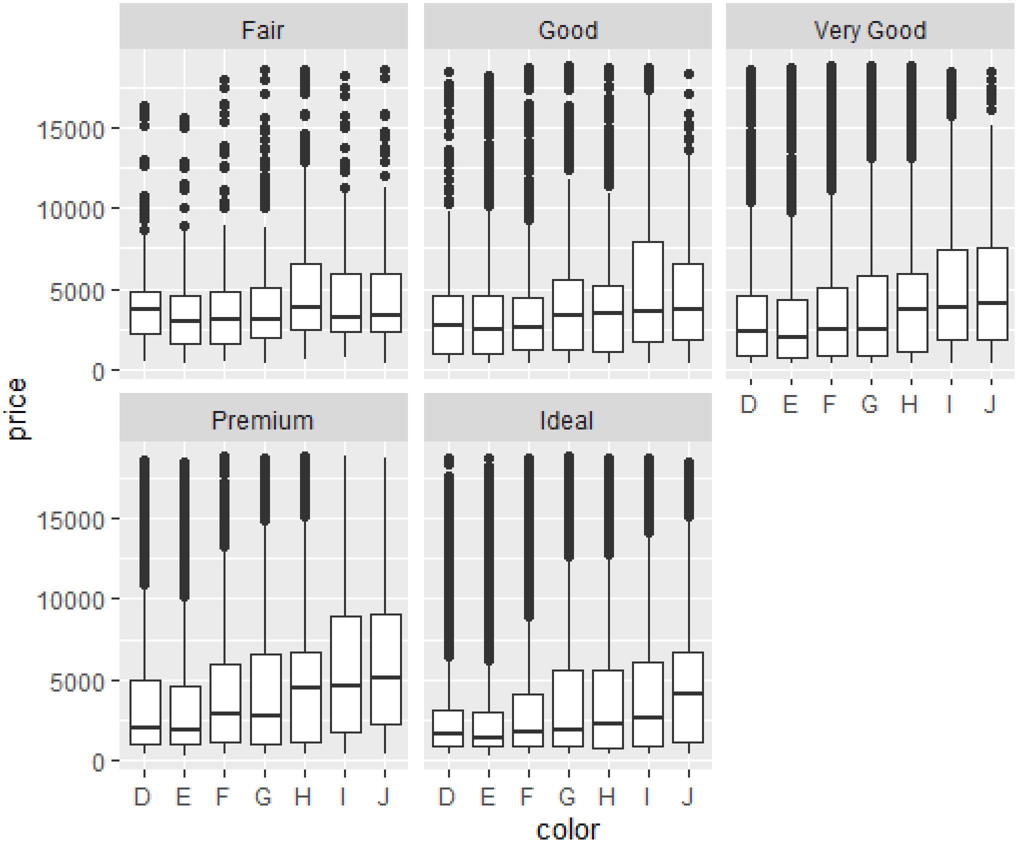
Rewrite the R code above so that Pclass is used in the aes and Gender is used in facet_wrap. Does the resulting plot change your conclusions at all?

(10) Finally consider how Survival is related to Age, Gender and Pclass.

```
# Survived by Age and Gender (beware colours)
ggplot(titanic, aes(x=Age)) +
  geom_histogram(aes(fill=Survived),binwidth=5) +
  facet_grid(Gender~Pclass)
```

Does this graphical plot change your conclusions at all?

Exercise. The "diamonds" dataset contains categorical variables *cut*, *color* and *clarity*. Write R code to reproduce the graphical plot below. Write down what you notice about the median price.



Summary

In this minilab, we have seen that when looking at relationships between categorical variables, exploratory data analysis often involves different types of bargraphs and summary tables of counts.