1. A counselor at ABC University wanted to find out the amount of outstanding tuition (in CAD$) ABC University international students carry in the Spring semester. A random sample of 50 ABC University international students was drawn in March to investigate this.
   Who: ABC university international students
   What: the amount of outstanding tuition (in CAD$)
   When: Spring semester
   Where: ABC University
   a) Identify the researcher. **[1 mark]**
      The counselor

   b) Provide a description of the objective. **[1 mark]**
      To find out the amount of outstanding tuition (in CAD$) ABC University international students carry in the Spring semester

   c) Identify the subjects of interest. **[2 marks]**
      Note: Make sure you also include the when and where, if available.
      ABC University international students in the Spring semester

   d) Provide a description of the variable of interest. Please also provide the units of measurement if it is a numerical variable or list out all possible classes if it is a categorical variable. **[1+1 marks]**
      Variable: amount of outstanding tuition
      Units: CAD$

   e) Identify the type of the variable and the corresponding scale of measurement. **[1+1 marks]**
      Note: Marks will be deducted with missing words (like "variable" or "scale").
      Type: numerical variable; Scale: interval/ratio scale

   f) Provide a description of the population of interest. **[1 mark]**
      All ABC University international students in the Spring semester

   g) Identify the most appropriate sampling method. **[1 mark]**
      Simple random sampling method

   h) Provide a description of the sample. **[1 mark]**
      Fifty randomly selected ABC University international students in March

   i) Is there an issue with the selection bias? Briefly justify your answer by comparing the (target) population and the sampling frame. **[0+2 marks]**
      Sampling frame = all ABC University international students in March
      The sampling frame does not match the target population perfectly, but March is part of Spring semester. So, the issue with the selection bias is not too bad.

2. Marisol lives in the City of Vancouver. She was planning to buy an electric vehicle (EV) this summer and wondering what percentage of EV owners have installed a Level 3 Supercharger at home. A sample of 20 EV owner living in her neighbourhood was drawn in June to investigate this.
   Who: EV owners (not EV themselves)
   What: whether the EV owners have installed a Level 3 Supercharger at home
   When: this summer
   Where: City of Vancouver
   a) Identify the researcher. **[1 mark]**
      Marisol

   b) Provide a description of the objective. **[1 mark]**
      To find the percentage of EV owners have installed a Level 3 Supercharger at home this summer

   c) Identify the subjects of interest. **[2 marks]**
      Note: Make sure you also include the when and where, if available.
      EV owners in the city of Vancouver this summer
      Note: If you only had "electric vehicles", it is wrong. Please make sure you understand why.

   d) Provide a description of the variable of interest. Please also provide the units of measurement if it is a numerical variable or list out all possible classes if it is a categorical variable. **[1+1 marks]**
      Variable: whether or not the EV owners have installed a Level 3 Supercharger at home
      Classes: Yes, No

   e) Identify the type of the variable and the corresponding scale of measurement. **[1+1 marks]**
      Note: Marks will be deducted with missing words (like "variable" or "scale").
      Type: categorical variable; Scale: nominal scale

   f) Provide a description of the population of interest. **[1 mark]**
      All electric vehicle owners in the City of Vancouver this summer

   g) Identify the most appropriate sampling method. **[1 mark]**
      Convenience sampling method

   h) Provide a description of the sample. **[1 mark]**
      Twenty EV owners living in her neighbourhood in the month of June

   i) Is there an issue with the selection bias? Briefly justify your answer by comparing the (target) population and the sampling frame. **[0+2 marks]**
      Sampling frame = all EV owners living in her neighbourhood in the month of June
      Since the sampling frame does not match the target population, there is an issue with the selection bias.

3. The director overseeing all senior homes in the Fraser Health Authority (FHA) wanted to know how many falls seniors have in 2023 that leads to major hip or lower body surgery. Ten seniors were randomly selected from each senior home in FHA region to form the sample.
   Note: This is a tricky question. Make sure you spend some time thinking about how to differentiate between a categorical variable from a numerical variable.
   Who: seniors
   What: number of falls that leads to major hip or lower body surgery
   When: 2023
   Where: senior homes in the FHA region
   a) Identify the researcher. **[1 mark]**
      The director of the FHA region

   b) Provide a description of the objective. **[1 mark]**
      To know how many falls seniors in the FHA region have that leads to major hip or lower body surgery in 2023

   c) Identify the subjects of interest. **[2 marks]**
      Note: Make sure you also include the when and where, if available.
      Seniors living in senior homes under FHA region in 2023

   d) Provide a description of the variable of interest. Please also provide the units of measurement if it is a numerical variable or list out all possible classes if it is a categorical variable. **[1+1 marks]**
      Variable: number of falls that leads to major hip or lower body surgery in 2023
      Units: no units (just counts)

   e) Identify the type of the variable and the corresponding scale of measurement. **[1+1 marks]**
      Note: Marks will be deducted with missing words (like "variable" or "scale").
      Note: I suggest you provide some reasoning to justify your choice of the type of variable.
      Type: categorical variable (as it is not likely to have 10 distinct outcomes)
      Scale: ordinal scale

   f) Provide a description of the population of interest. **[1 mark]**
      All seniors living in senior homes under FHA region in 2023

   g) Identify the most appropriate sampling method. **[1 mark]**
      Stratified random sampling method

4. Students are generally confused with the stratified random sampling method and the cluster sampling method. I hope students have a better understanding of them after completing this question.
   There are two situations in the questions: one method is used in each of the two situations. In each of the following two situations, (1) identify the sampling method used, (2) describe how you would define the non-overlapping groups, and (3) outline the three main steps.
   Note: In practice, geographic factor should not be used to do the stratification. It is used here only for the sake of illustration.
   a) To get an idea of how much detached houses in the city of Vancouver cost these days, Joe went to a real estate web site to collect some information. He divided the city of Vancouver into 20 communities; randomly selected 5 detached houses from each community; and the 100 (20 communities times 5 houses per community) detached houses formed the sample. **[1+1+3 marks]**
      1) Method: Stratified random sampling method [1]
      2) Subjects: detached houses; Demographics = 20 areas/communities in the city of Vancouver [1]
      3) Three main steps:
      a) All detached houses in Vancouver are divided in to 20 non-overlapping communities or strata. [1]
      b) Five detached houses (or subjects) are randomly selected from each community. [1]
      c) The 20 mini random samples of five houses become the stratified random sample. [1]

   b) To get an idea of how much detached houses in the city of Vancouver cost these days, Joe went to a real estate web site to collect some information. He divided the city of Vancouver into 20 communities; randomly selected 5 of the communities; and all detached houses (that are listed on the web site) within those 5 communities were chosen in the sample. **[1+1+3 marks]**
      1) Method: Cluster sampling method [1]
      2) Subjects: detached houses; Clusters = 20 areas/communities in the city of Vancouver [1]
      3) Three main steps:
      a) All detached houses in the city of Vancouver are divided in to 20 non-overlapping communities or clusters. [1]
      b) A random subset of 5 communities is selected from all 20 communities. [1]
      c) All detached houses in the 5 selected communities (listed on the website) become the cluster sample. [1]

5. Consider a movie theatre with 30 rows of 20 seats in each row. There are some prize giveaways before the movie starts. Identify the most appropriate sampling method used below.
   Note: In practice, geographic factor should not be used to do the stratification. It is used here only for the sake of illustration.
   a) Two random rows will be drawn (from among all 30 rows). Everybody in the two selected rows (40 of them) will get a prize. **[1 mark]**
      Cluster Sampling Method (with random selection of rows and everybody in the rows get the prize)

   b) Each person entering the theatre will have to write down their names on a piece of paper and then put the paper in a big bag. Twenty (20) names were randomly drawn for the prize. **[1 mark]**
      Simple Random Sampling Method

   c) Two random winners will be drawn in each row, for all 30 rows. **[1 mark]**
      Stratified Random Sampling Method (all rows are used and random sample in each row)

   d) The first 30 movie goers who enter the theatre get a prize. **[1 mark]**
      Convenience Sampling Method

   e) A random person is chosen among the first five who enter the theatre and prizes are given to every 5^{th} person thereafter. **[1 mark]**
      Systematic Sampling Method

6. General public underestimate the power of statistics. In particular, misuse of statistics could have a devastating impact on individuals or organizations.
   Your task is to do a Google search about the 1936 US Presidential Election between the incumbent Democratic candidate Franklin D. Roosevelt and Republican candidate Alf Landon. You want to pay special attention to the pre-election prediction between the two organizations – *Literary Digest* and *Gallup Poll*. Answer the following questions.
   Note: You might get slightly different values from different websites. So, there are no standard answers here.
   a) What was the success rate (predicting the correct election outcome) of *Literary Digest* prior to 1936? **[1 mark]**
      The Literary Digest had been predicting the winners of US Presidential Election since 1916 (5 elections in a row).

   b) What were the 1936 pre-election prediction by both *Literary Digest* and *Gallup Poll*? Please focus on your answer as 1) who would win the 1936 Presidential election and by what percentage of popularity vote. **[2 marks]**
      Most sources say that the *Literary Digest* had predicted the Republican candidate Alf Landon would win the election and get 57% of the popularity vote.
      Most sources say that the Gallup Poll had predicted the Democratic candidate Franklin Roosevelt would win the election and get about 56% of the popularity vote.

c) What was the official result of the 1936 US Presidential Election? **[1 mark]**
Most sources say that the official result that Roosevelt got about 62% of the popularity vote.

Now let us focus on the *Literary Digest* only from this point on.

d) In relation to selection bias, what did the *Literary Digest* do (or not do) to wrongly predict the election results? Please provide as much details as possible. **[2 marks]**
In short, the *Literary Digest* had selected the wrong group of subjects. Here, the target population is supposed to be all American people (or registered voters). The sampling frame is all names that appeared in phone books and on club membership lists. Note that at the time, only the rich person could afford telephones and club membership.

e) In relation to non-response bias, what did the *Literary Digest* do (or not do) to wrongly predict the election results? Please provide as much details as possible. **[2 marks]**
The Literary Digest sent out about 10 million in regular mail, but only about 2.4 million returned. So, only about 24% of the survey were returned. Hence, there was a serious non-response bias. Also, those who responded were typically more vocal than those who did not respond.

f) In relation to response bias, what did the *Literary Digest* do (or not do) to wrongly predict the election results? Note that The Great Depression started in 1929 and last till around 1939. So, general public did not know when the Great Depression would end. So, imagine to whom you would point your fingers when time is tough. **[2 marks]**
With The Great Depression as the backdrop, it is human nature that general public would tend to point fingers to the incumbent government (the Democrats with Roosevelt as the president) when times were tough. As a result, the typical Democratic supporters might say that they would vote for the Republican to the pollsters but ended up still voting for the Democrats.

g) Now focus on modern days. What are your thoughts about 2016 US Presidential Election between Democratic candidate Hilary Clinton and Republican candidate Donald Trump, in relation to the three biases that we have learned? **[2 marks]**
Here is my take.
1) Selection bias is greatly reduced.
2) Non-response bias is always there (typically 30-35% response rate), but it has been kept at a level where it did not have a big impact on the results, especially when the stratification is done well.
3) Response bias is still the main issue in practice. It is something that we have no control over no matter how careful the sampling and data collection are administered. And the response bias is also what we all should be careful of when handling data.

7. A police officer from Vancouver Police Department wanted to find out the percentage of drivers who were distracted (defined as using their phone while driving or waiting at the traffic lights) during the day. The officer took a random sample of 80 drivers to investigate this. The results can be found in "DANA4800_HW1_Q7_Data.xlsx" on BrightSpace.

   a) Create a Frequency Table of the variable "Distracted" using the *table()* function. **[1 mark]**
   Note: When copy-and-pasting text output from R to Word document, for example, make sure you use "fixed-width fonts", like **Courier New**. Otherwise, the output does not look right or aligned properly.

```
Distracted
 No Yes
 36  44
```

   b) Create a Probability Table of the variable "Distracted" using the *proportions()* function. **[1 mark]**

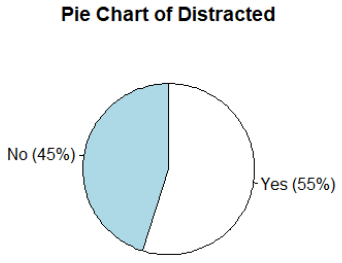
```
Distracted
  No  Yes
0.45 0.55
```

   c) Provide a description of the parameter of interest. **[2 marks]**
   Note that the objective is to find the percentage of drivers who are distracted. Therefore, the answer here should be:
   The proportion (or percentage) of all drivers who are distracted when driving during the day. [2]

   d) Provide a description of the corresponding statistic. **[2 marks]**
   The proportion (or percentage) of the 80 randomly selected drivers who are distracted when driving during the day. [2]

   e) Calculate the value of the most appropriate statistic. **[1 mark]**
   $\bar{p} = \frac{p}{n} = \frac{44}{80} = 0.55$ **or 55%**

   f) Produce a Pie Chart using the *pie()* function, with the "clockwise" arguments set as TRUE, and "Yes" goes before "No". Please also submit the code to produce such graph. Please use fixed-width fonts. **[2+1 marks]**
   Note: Please make sure you personalize the pie chart by including the Main Title, and add labels to axes (if applicable) etc.

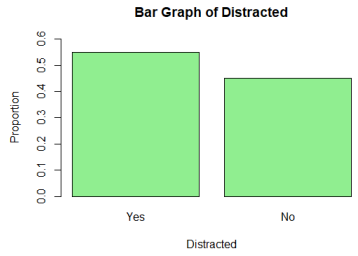


[2]

```
counts <- counts[c("Yes","No")]
pie(counts,
    main = "Pie Chart of Distracted",
    clockwise = TRUE,
    labels = c("Yes (55%)","No (45%)"))
```

g) Provide a description of the above pie chart. **[1 mark]**
Majority of the drivers are distracted in the sample of 80 drivers. [1]
Note: Any reasonable answers are fine, but it has to be describing the graph.

h) Produce a Bar Graph using the *barplot()* function, "Yes" goes to the left of "No". Please also submit the code to produce such graph. Please use fixed-width fonts. **[2+1 marks]**
Note: A graph directly copied from Excel without any annotation will get a zero.



[2]

```
counts <- counts[c("Yes","No")]
barplot(props,
        main = "Bar Graph of Distracted",
        xlab = "Distracted",
        ylab = "Proportion",
        ylim = c(0,0.6),
        col = "light green")
```

i) Which graph is better to use here? Briefly justify your answer using statistical reasoning. **[0+1 mark]**
Either graph is fine because the data is in nominal scale. [1]

Question #8 is meant to be for you to practice the calculations manually. That said, feel free to use R (or Excel) to double-check your work before submission.

8. Whenever there is a major concert in a city, the hotel rate during that time normally go up. A random sample of 20 hotels in downtown Vancouver was drawn during the time of major concert and the rate of a hotel room per night (based on two double-bed rooms) was recorded.

286	378	245	292	244	314	298	282	281	317
319	237	289	275	285	227	270	322	274	293

a) Identify the subjects of interest. **[1 mark]**
Subjects: hotels in downtown Vancouver during major concert time

b) Calculate the average hotel room rate manually. Please show all work. **[1 mark]**
$\overline{X} = 286.4$ **or $286.4 [2]**
Note: Work is not shown here but feel free to let me know if you have trouble doing it.

c) Provide a description of the statistic in part (b). **[2 marks]**
The average hotel room rate per night of the 20 randomly selected hotels in downtown Vancouver during major concert time. [2]

d) Find the median hotel room rate manually. Please show all work. **[2 marks]**
First, the observed values have to be arranged in ascending order.

227	237	244	245	270	274	275	281	282	285
286	289	292	293	298	314	317	319	322	378

There are 20 observed values. Therefore, the location of median is $\boldsymbol{i = \frac{20+1}{2} = 10.5}$. In other words, median $\boldsymbol{= \frac{285+286}{2} = 285.5}$. [2]

e) Use the method in our notes to find the first quartile and the third quartile. Then find the interquartile range. **[1+1+1 marks]**
$\boldsymbol{Q_1}$ is the median to the left of the overall median 285.5. In other words, it is the midpoint between 270 and 274. Therefore, $\boldsymbol{Q_1 = \frac{270+274}{2} = 272}$. [1]
$\boldsymbol{Q_3}$ is the median to the right of the overall median 285.5. In other words, it is the midpoint between 298 and 314. Therefore, $\boldsymbol{Q_3 = \frac{298+314}{2} = 306}$. [1]
Interquartile range = 306 – 272 = 34. [1]

f) Determine if there is/are any outlier(s), using IQR. **[3 marks]**
$\boldsymbol{LL = 272 - 1.5 \times 34 = 272 - 51 = 221}$ [0.5]
There is no outlier on the left-hand side because the minimum (227) is not less than the lower limit of 221. [1]
$\boldsymbol{UL = 306 + 1.5 \times 34 = 306 + 51 = 357}$ [0.5]
There is one outlier on the right-hand side because maximum (378) is larger than the upper limit of 357. [1]

The following two parts require the use of R.

g) Use the *summary()* function to find the five-number summary of the hotel rates. **[1 mark]**

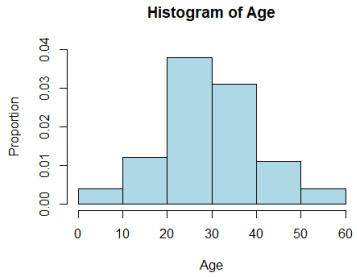
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  227.0   273.0   285.5   286.4   302.0   378.0
```

9. A first-year Langara student wanted to know the weekly expenses (in CAD$) of typical Langara students. To investigate this, she got a random sample of 100 students this term. The data set is in "Expense" worksheet of the file "DANA4800_HW1_Q9_Data.xlsx" on BrightSpace.

a) Use the *summary()* function to find the Five-Number Summary. **[1 mark]**

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   3.00   23.75   29.50   29.90   36.25   58.00
```

b) Use the *hist()* functions to produce a Histogram with proportion on the y-axis with 6 bars only. Specifically, please make sure there are 6 bars (1-10, 11-20, …, 51-60), chart title included and axes labelled properly. **[4 marks]**
Note: Make sure you label your graph and axes appropriately.

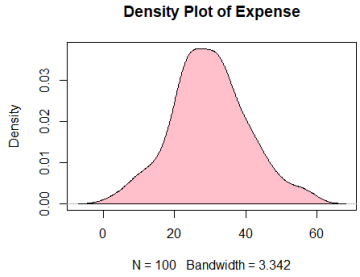


[2]

c) Provide a description of the above graph. **[2 marks]**
Centre: between $20 and $30
Spread: $0 to $60
Shape: slightly skewed to the right
Outlier: No obvious outliers

d) Use the following segment of code to produce a Density Curve. Provide a description of the shape (only). **[2+1 marks]**
Note: Make sure you label your graph and axes appropriately.

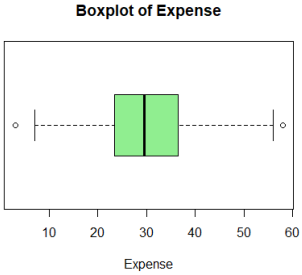
```
dens <- density(INSERT YOUR VARIABLE HERE)
plot(dens)
polygon(dens, col = "pink")
```



[2]

Shape = pretty much symmetrical or even Normal [1]

e) Use the *boxplot()* function to make a horizontal Boxplot. **[2 marks]**
Note: Make sure you label your graph and axes appropriately.



[2]