**Name (*Last, First*): Joanna Lu, ID: 300916162**

**Instructions**:

1) You can provide complete solutions in the space provided or screenshot your answers and attach them to this assignment.

2) This assignment has two parts. Make sure you follow all given instructions in the project guide.

***3) Late assignments will not be accepted. Due date is:***

**Overview**

This project should provide an opportunity for you to demonstrate your proficiency in:

1. Collecting, finding, or producing data (Part 1 of Assignment)

2. Performing Exploratory Data Analysis (Part 1 of Assignment)

3. Performing hypothesis test (Part 2 of Assignment)

Develop a question of interest to you that will be answered by performing the 3 tasks listed above.

**Part 1: Mini Research project #1 (2%, 20 marks)**

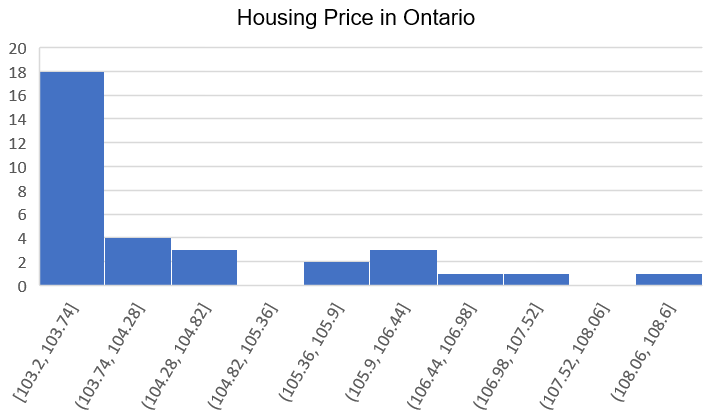
**Formulating the question and obtaining the data**

**Step 1: Format the question and collect Data**

1. This data is retrieved from Statistics Canada. It is about the housing price in Ontario from January 2018 to September 2020. (Focus on housing price only, highlighted in yellow color). Based on the statistics, the mode is 103.5; it shows the most frequently occurring housing price is around $1,035,000. Does the average housing price more than or less than $1,035,000?



Source: <https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1810020501&pickMembers%5B0%5D=1.17&cubeTimeFrame.startMonth=01&cubeTimeFrame.startYear=2018&cubeTimeFrame.endMonth=11&cubeTimeFrame.endYear=2020&referencePeriods=20180101%2C20201101>

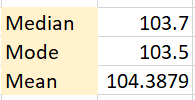
**Step 2: Graphs and charts**

x: housing price (ten thousand of $)

**Step 3: Descriptive statistics (summary statistics)**

1. **Calculate measures of central tendency**

* Formula of each item is attached in excel.

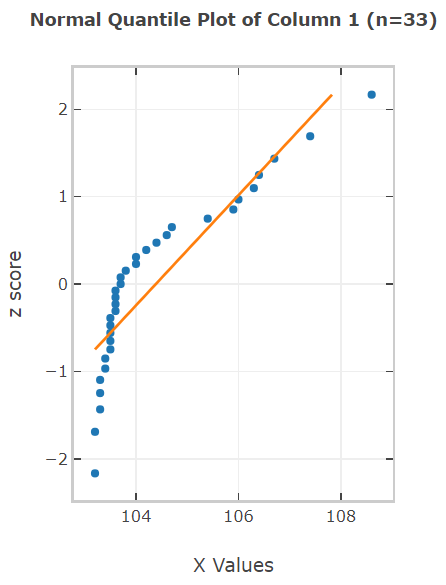
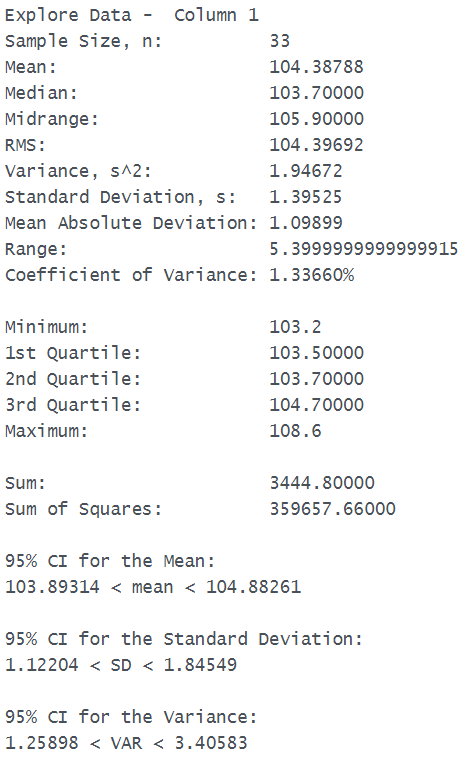




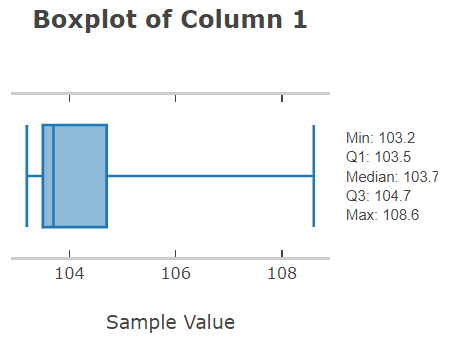
1. **Calculate measures of variation**

* Sample variance: 1.94672
* Sample Standard Deviation: 1.39525

1. **Exploratory data analysis**



* From the graph above we can see there are some outliers on the both sides.
* **Boxplot** shows the minimum, maximum, Q1, Q2 (Median), and Q3 number of housing price.
* The lowest housing price is $1,032,000.
* The highest housing price is $1,086,000.



**Step 4: Analysis and Results**

* The median is the data which located at the middle of all collected data. In this case, the data size is 33, so the median data would be 17th data. That is $1,037,000.
* The mode shows the most frequently occurring housing price is around $1,035,000, and the mean shows the average housing price is around $1,043,879 which is adding all data together, then divide the number of data.
* According to the histogram, the shape of the distribution is right skewed.
* The calculation below indicates what the mild outlier and extreme outlier are.
* Outlier:

IQR = Q3 – Q1 = 104.7 – 103.5 = 1.2

* mild outlier:

Q1 – 1.5IQR = 103.5 – 1.5(1.2) = 101.7 (no data less than 101.7)

Q3 + 1.5IQR = 104.7 + 1.5(1.2) = 106.5 (106.7, 107.4, 108.6 greater than 106.5, so they are mild outlier)

* extreme outlier:

Q1 – 3IQR = 103.5 – 3(1.2) = 99.9 (no data less than 99.9)

Q3 + 3IQR = 104.7 + 3(1.2) = 108.3 (108.6 > 108.3, so 108.6 is extreme outlier.

* The extreme outlier is 108.3, which means if the housing price over $1,083,000, then it may be considered as overpricing from buyers’ point of view, however, it will be a good deal for sellers.

In addition, the mild outlier is 106.5, so for the housing price over $1,065,000 that are slightly expensive than average. In this case, $1,067,000 and $1,074,000 can be treated as mild outliers.

* Standard deviation shows the gaps between data. In addition, variation is sensitive to outliers. If we include all data for calculating the standard deviation, the result will bigger than the one excludes outlier.
* In conclusion, the average housing price is $1,043,879; it is a little bit higher than the mode price $1,035,000.

**Part 2: Mini Research project #2 (3%)**

**Step 1: Format the question and collect Data**

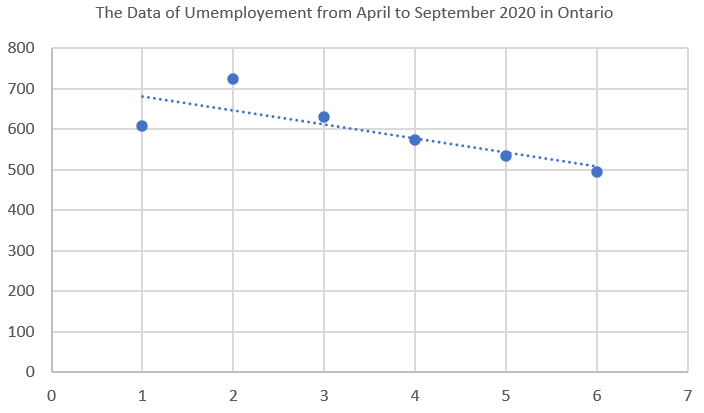
1. This data is retrieved from Statistics Canada about the population of unemployment, age above 25 years old in Ontario. Time duration is from April 2020 to September 2020. April is one of the peak months of COVID-19 where most employees are laid off during the time. Does the unemployment in September higher than unemployment in April?

**Source:** <https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1410028703&pickMembers%5B0%5D=3.1&pickMembers%5B1%5D=4.5&cubeTimeFrame.startMonth=05&cubeTimeFrame.startYear=2020&referencePeriods=20200501%2C20200501>

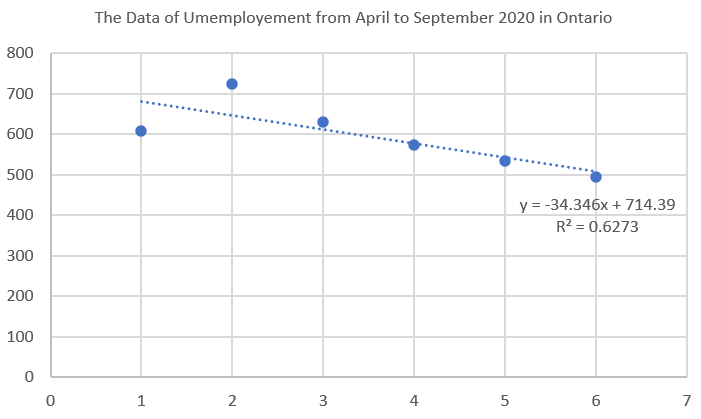
* Data:

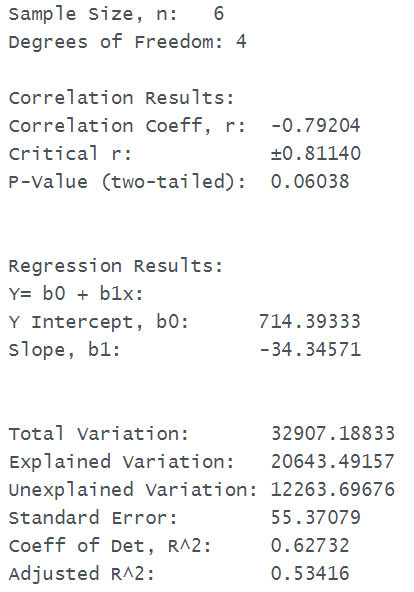


**Step 2: Graphs and the line of best fit**

* Make a scatter plot and draw a line of best fit using statistical software.
* x: month (From April 2020 to September 2020)
* 1 represents April 2020, 2 represents May 2020, …, 6 represents September 2020
* y: population of unemployment in thousands

**Step 3: Descriptive statistics (summary statistics)**

1. Find the equation of a line of best fit
2. Find the correlation coefficient and coefficient of determination.



**Step 4: Analysis and Results**

* Discussion on the relationship between chosen variables (positive/negative, strong/moderate/weak/none) and draw conclusions from findings.

The correlation coefficient, which is r = – 0.792; it is around – 0.8, so we can consider that it is a negative and moderate relationship between the month and the unemployment.

In summary, from the scatter diagram we can see, the unemployment is reducing due to the negative value of correlation coefficient (r = – 0.792). For example, the unemployment in September is lower than the unemployment in April.

September:

* x-axis: 6
* y-axis: 500

April:

* x-axis: 1
* y-axis: 600