Statistical Analysis using RStudio on Research Project Investments in Toronto, Canada

University of Toronto | Yuze Fu, Fanglu Yang, Michael Stevenson Ong

2024-07-20

1. Rationale

1.1 Summary of Data

```
## Rows: 5,521
## Columns: 22
## $ X_id                     <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 1~
## $ Program                  <chr> "Ontario Research Fund - Research Infrastruc~
## $ Project.Number           <chr> "42773", "42799", "42789", "43525", "43535",~
## $ Project.Title            <chr> "Toronto High Containment Facility (THCF)", ~
## $ Project.Description      <chr> "The Toronto High Containment Facility is a ~
## $ Area.Primary             <chr> "RDF30", "RDF30", "RDF30", "RDF20-21", "RDF2~
## $ Area.Secondary           <chr> "RDF301", "RDF301", "RDF301", "RDF21001", "R~
## $ Discipline.Primary       <chr> "N/A", "N/A", "N/A", "N/A", "N/A", "N/A", "N~
## $ Discipline.Secondary     <chr> "N/A", "N/A", "N/A", "N/A", "N/A", "N/A", "N~
## $ Round                    <chr> "Biosciences Research Infrastructure Fund", ~
## $ Approval.Date            <chr> "2023-04-28T00:00:00", "2023-04-28T00:00:00"~
## $ Lead.Research.Institution <chr> "University of Toronto", "McMaster Universit~
## $ City                     <chr> "Toronto", "Hamilton", "London", "Sarnia", "~
## $ Ontario.Commitment       <chr> "$9,931,992.00 ", "$2,157,853.00 ", "$3,910,~
## $ Total.Project.Costs      <chr> "$109,516,266.00 ", "$14,382,498.00 ", "$38,~
## $ Keyword                  <chr> "Virology, microbiology, emerging pathogens,~
## $ EXPENDITURE_TYPE         <chr> "Capital", "Capital", "Capital", "Capital", ~
## $ Salutation               <chr> "Dr.", "Dr.", "Dr.", "Dr.", "Dr.", "Dr.", "D~
## $ First.Name               <chr> "Scott", "Matthew ", "Eric", "Mehdi", "Mehdi~
## $ Middle.Name              <chr> "", "", "", "M.", "M.", "M.", "N.", "G.", "F~
## $ Last.Name                <chr> "Owen", "Miller", "Arts", "Sheikhzadeh", "Sh~
## $ OIA.AREA                 <chr> "N/A", "N/A", "N/A", "N/A", "N/A", "N/A", "N~
```

The dataset is retrieved from the Ontario Data Catalogue from https://data.ontario.ca/dataset/ontario-research-funding-summary that was last refreshed in March 18, 2024 regarding comprehensive overview of research projects funded by the The Ministry of Colleges and Universities, from the year 2004 to 2024, in Ontario. This dataset has a total of 5511 reported entries with 21 variables.

The following is a breakdown of the variables that this report uses:

1. Approval.Date = (chr) The date when the research funding is approved
2. Land.Research.Institution = (chr) The lead institution doing the research

3. City = (chr) The city where the institution is located
4. Ontario.Commitment = (chr) The amount of money in Canadian Dollars that The Ministry of Colleges and Universities is funding
5. Total.Project.Costs = (chr) The amount of money in Canadian Dollars that the actual project costs

1.2 Background of Data

The Ministry of Colleges and Universities stores a detailed overview of research projects in Ontario including the program type, project title and its description, approved date, lead research institution and its city, the amount of funding provided, the actual cost of the project, and other relevant administrative information. This dataset which is updated annually, contains all government-recognized research projects between October 27, 2004 and March 31, 2024.

1.3 Data Cleaning

```
##                         Lead.Research.Institution year     city commitment
## 1                             University of Toronto 2023  Toronto    9931992
## 2                               McMaster University 2023 Hamilton    2157853
## 3                                Western University 2023   London    3910155
## 4  Lambton College of Applied Arts and Technology 2023   Sarnia     702470
## 5  Lambton College of Applied Arts and Technology 2023   Sarnia     998759
## 6  Lambton College of Applied Arts and Technology 2023   Sarnia     950436
## 7 Fanshawe College of Applied Arts and Technology 2023   London     353343
## 8                              University of Ottawa 2023   Ottawa    4986999
##        cost
## 1 109516266
## 2  14382498
## 3  38873876
## 4   1756830
## 5   2499854
## 6   2376781
## 7    883597
## 8  12467498
```

Some data cleaning is done (**dclean**) to fix improper datatype, such as the *Ontario.Commitment* and *Total.Project.Costs* are made *num* instead of *chr*, without the dollar ($) and comma sign, as well as to cater the need of extracting the 'year' only out of *Approval.Date* where it is made *num* instead of *chr*.

1.4 Research Question

How does research project investments by Ministry of Colleges and Universities vary accross different colleges and universities in Toronto?

2. Analysis

2.1 General Analysis

```
## # A tibble: 6 x 3
##    year commitment       cost
```

```
##   <dbl>      <dbl>      <dbl>
## 1  2004   55442115  486464662
## 2  2005   56211712  140129592
## 3  2006  147334888  470828738
## 4  2007  300962363  835501117
## 5  2008   46051213  101431738
## 6  2009  367086216 1076915327
```

data1 shows Ministry of Colleges and Universities' investments on research projects through the years. Unfortunately, it suggests no obvious trend as it is rather random year per year. The following sub-sections will show other analyses.

2.2 How the Research Question is Formed

```
## # A tibble: 8 x 3
##   city     commitment        cost
##   <chr>         <dbl>       <dbl>
## 1 Toronto  1222457264 3960978514.
## 2 London    305669816  908896167
## 3 Ottawa    286555282  904477262
## 4 Waterloo  275228726  852629861.
## 5 Hamilton  259697482  816687909.
## 6 Kingston  257175282  861343518
## 7 Guelph    134908933  459337071
## 8 Windsor    22569621   62394330
```

data2 orders the research projects with the most *commitment* in comparison to cities in Ontario. Since Toronto holds the most amount of money that the Ministry of Colleges and Universities invest on and the total cost of the research, this report will be narrowed down to colleges and universities in Toronto only.

2.3 Findings for Colleges and Universities in Toronto only.

```
## # A tibble: 8 x 4
##   institution                                              N commitment   cost
##   <chr>                                                <int>      <dbl>  <dbl>
## 1 University of Toronto                                 1147  581466780 1.96e9
## 2 York University                                        200   52672305 1.70e8
## 3 Toronto Metropolitan University (formerly Ryerson Uni~  115   27691631 7.83e7
## 4 Ontario College of Art and Design (OCAD) University     10    6695925 1.67e7
## 5 George Brown College of Applied Arts and Technology      7    6155903 2.03e7
## 6 Royal Military College of Canada                         4    1527261 3.85e6
## 7 Humber College                                           2    2000000 5.37e6
## 8 Seneca College of Applied Arts and Technology            2    1150000 2.89e6
```

data3 orders the research projects with the most investment from Ministry of Colleges and Universities (*commitment*) in comparison to the colleges and universities in Toronto. It suggests that University of Toronto takes a majority portion of the research projects in Toronto, given it has the most amount of research projects (N), the most investment from Ministry of Colleges and Universities (*commitment*), and the most actual project cost (*cost*).

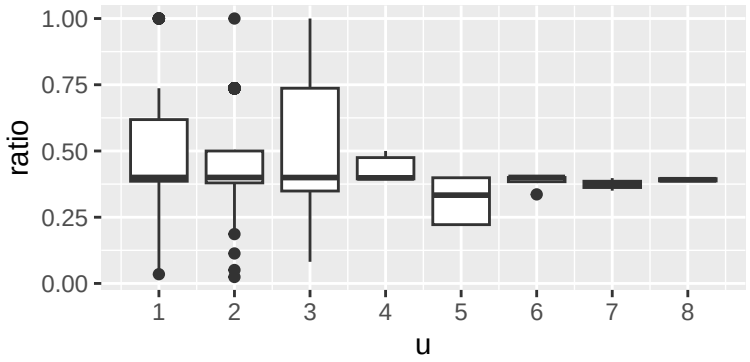
2.4 Averages of Sub-Section 2.3

```
## # A tibble: 8 x 5
##   institution                                N avg_commitment avg_cost avg_ratio
##   <chr>                                <int>          <dbl>    <dbl>     <dbl>
## 1 University of Toronto                 1147        506946. 1709036.     0.297
## 2 York University                        200        263362.  848796.     0.310
## 3 Toronto Metropolitan University (form~  115        240797.  680715.     0.354
## 4 Ontario College of Art and Design (OC~   10        669592. 1667598.     0.402
## 5 George Brown College of Applied Arts ~    7        879415. 2906391.     0.303
## 6 Royal Military College of Canada          4        381815.  963554.     0.396
## 7 Humber College                            2       1000000  2685066.     0.372
## 8 Seneca College of Applied Arts and Te~    2        575000  1446075      0.398
```

data4 modifies *commitment* and *cost* data from **data3** to each of their respective averages per research per college/university (*avg_commitment* and *avg_cost*). A new column *avg_ratio* is created to calculate the average proportion of investment by Ministry of Colleges and Universities on research projects relative to the average cost (i.e. avg_ratio = avg_commitment ÷ avg_cost. It turns out that University of Toronto with the most research projects receives the least investment, while other institutions with significantly less amount of research projects receive more investment.

3. Graphical Representations and Plots

3.1 Boxplot of Proportion of Investment-to-Cost per Project

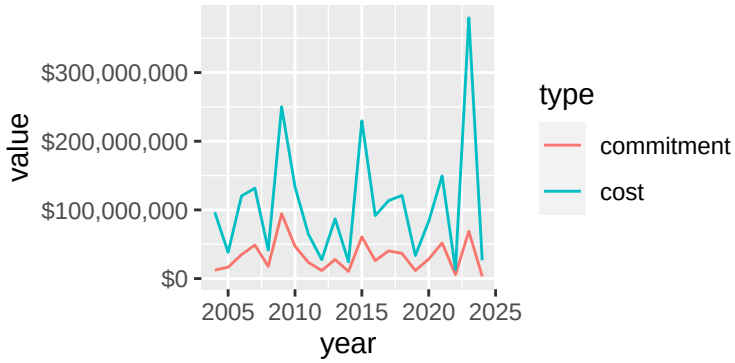


1 - University of Toronto
2 - York University
3 - Toronto Metropolitan University (formerly Ryerson University)
4 - Ontario College of Art and Design (OCAD) University
5 - George Brown College of Applied Arts and Technology
6 - Royal Military College of Canada
7 - Humber College
8 - Seneca College of Applied Arts and Technology

data5 is a boxplot that illustrates the proportion of Ministry of Colleges and Universities' investment (*commitment*) to the total cost (*cost*) per project as *ratio*, which explains why the value range is [0-1], where a ratio closer to 0 means the project is under-invested and a ratio closer to 1 means the project is highly-invested. York University is found to have the most outliers with more under-invested projects, while that and University of Toronto both have extreme outliers as shown by the plot on the endpoints. Since there are extreme outliers existing very far from the interquartile range, the normality of the data is questionable.

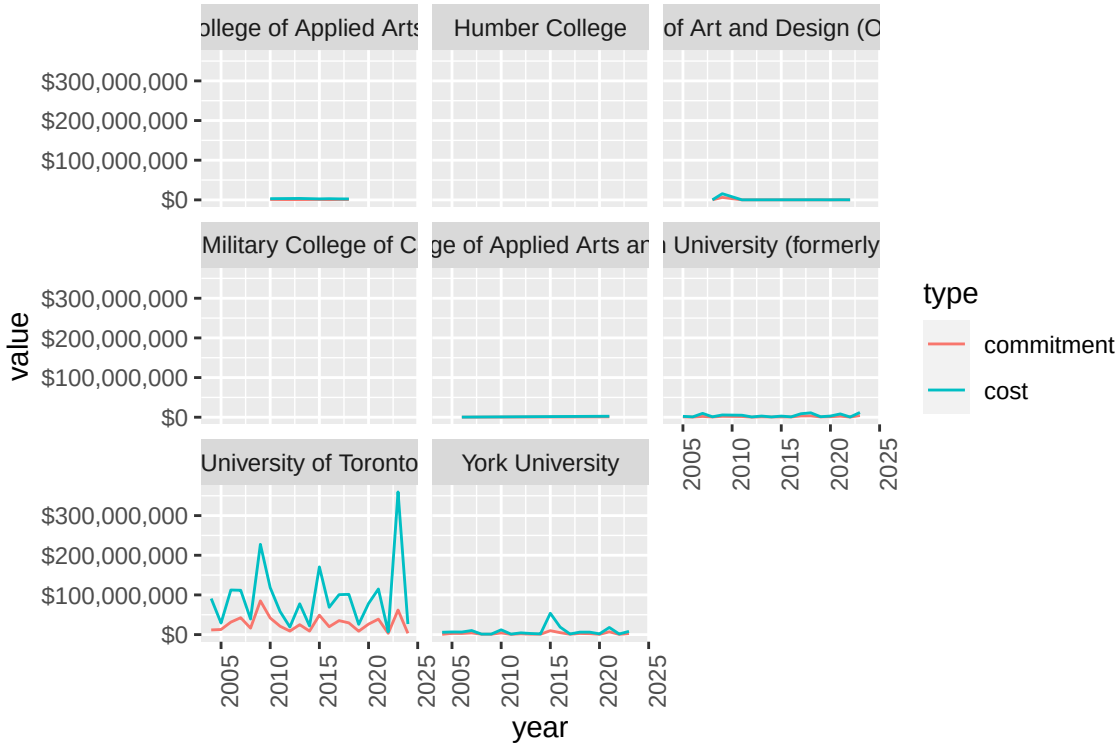
Therefore, using sampling techniques such as bootstrapping is recommended because it will not depend on the distribution of the data.

3.2 Line Graph of Investment and Cost per Project



data6 is a line graph that shows the trend of total investment by Ministry of Colleges and Universities (*commitment*) compared to the total cost (*cost*) for every research project per year between 2004 to 2024. Generally, the investments (*commitment*) follow the trend of the cost (*cost*) as seen from the similar "bends". However, when the cost spikes high, the investment does not follow as high. In other words, the investment (*commitment*) does not cover as much as the research project cost *cost* when it costs significantly higher.

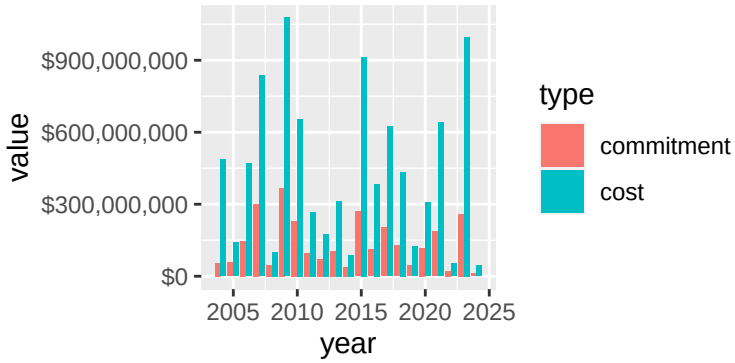
3.3 Line Graph of Investment and Cost per Project based on Universities



data7 narrows down **data6** further by dividing the cases to each colleges and universities in Toronto of the same year period (2004-2024). At a first glance, the University of Toronto seems to dominate both

investments from Ministry of Colleges and Universities (*commitment*) and total research project costs (*cost*) as other universities shows a small line towards the bottom of the graph. Moreover, the University of Toronto graph seems almost the same to **data6**, indicating that University of Toronto actually dominates.

3.4 Bar Graph of Investment and Cost per Project based on University of Toronto



data8 is a bar graph that portrays a detailed breakdown of total investment by Ministry of Colleges and Universities (*commitment*) and the total cost (*cost*) of every research project per year from 2004-2024 for University of Toronto only, the dominating institution in Toronto. The general observation is similar as the previous one, where the *commitment* fluctuation follows the *cost*. There are several research project cost that peak at 2003 and 2023, however 2023's investment (*commitment*) with higher *cost* is less than 2007's and 2015's despite the lower *cost*. To conclude statistically whether the govenment generally supports research project initiatives (financially) needs a modelling done in the next sections.

4. Using Confidence Interval and Test of Hypothesis

4.1 T-Test with Default Confidence Interval

```
##
##  One Sample t-test
##
## data:  t_t$ratio
## t = 110.9, df = 1992, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.4635237 0.4802132
## sample estimates:
## mean of x
## 0.4718684
```

A t-test code has been run and the range of the mean value of *ratio* (i.e., the proportion of investment by Ministry of Colleges and Universities to the actual research project cost) is between 0.4635237 and 0.4802132, with the mean of the data being 0.4718684

4.2 Bootstrap Confidence Interval

```
##      2.5%     97.5%
## 0.4636792 0.4804520
```

As concluded from the boxplot (sub-section 3.1) where the normality of the data is questionable, a bootstrap confidence interval test is done as shown above. After running the code several times, we observe that the bootstrapped data gives around the same mean value range of *ratio* by t-test as tested previously.

4.3 Proportion Test

```
##
##  1-sample proportions test with continuity correction
##
## data:  t_p$commitment out of t_p$cost, null probability 0.5
## X-squared = 580273283, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.3086107 0.3086395
## sample estimates:
##         p
## 0.3086251
```

The proportion test is given to the actual formula of *ratio* which is proportion of *commitment* to *cost*. Since the p-value $<$ 2.2e-16, there is a strong evidence that the proportion of commitment to cost is significantly different from 0.5. The sample proportion of *commitment* is 0.3086251 with a 95% confidence interval between the range 0.3086107 and 0.3086395. In other words, there is 95% confidence that the true proportion of *commitment* (to *cost*) falls within the said range.

4.4 Hypothesis Test

```
##
##  One Sample t-test
##
## data:  d$ratio
## t = -13.058, df = 5520, p-value = 1
## alternative hypothesis: true mean is greater than 0.5
## 95 percent confidence interval:
##  0.462765      Inf
## sample estimates:
## mean of x
## 0.4669311
```

Hypothesis Testing is done to test whether or not the average *ratio* is greater than 0.5, as another effort to strengthen the previous findings. Looking at the p-value which is much greater than $\alpha = 0.1$, there is no sufficient evidence to reject the null hypothesis, therefore it is concluded that the average *ratio* is equal to or smaller than 0.5.

4.5 Equality of Variance Test

```
##   group_ind group_mean    group_var
## 1        G1   506945.8 1.374531e+12
## 2        G2   613375.4 1.902099e+12
```

```
##
##  F test to compare two variances
```

```
##
## data:  RV by group_ind
## F = 0.72264, num df = 1146, denom df = 1992, p-value = 1
## alternative hypothesis: true ratio of variances is greater than 1
## 95 percent confidence interval:
##  0.6632934       Inf
## sample estimates:
## ratio of variances
##          0.7226393
```

Variance Test is done to compare two populations, where the first population is the research projects done by University of Toronto and the second population is the research projects done by all colleges and universities in Toronto. This is done because sub-section 3.3 concludes that University of Toronto is the dominating institution in both *commitment* and *cost*. The code is run and it shows that the ratio variance of *commitment* of University of Toronto towards all colleges and universities in Toronto is 0.7226393, which means University of Toronto has less variance. The p-value which is much greater than $\alpha = 0.1$ also shows that there is no sufficient evidence to reject the null hypothesis, therefore it is concluded that the ratio of the variances of *commitment* between the two populations is equal to or smaller than 1. The following conclusion can be taken: University of Toronto which seems to have the most research projects in Toronto has less fluctuation in *commitment*, while other universities with fewer research projects have more fluctuation in *commitment*, i.e. they receive so little in some projects and receive so large in other projects.

5. Regression Models

5.1 Linear Regression of *commitment* to *cost*

```
##
## Call:
## lm(formula = d$commitment ~ d$cost, data = d)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -23185037    -96863    -48751    -32662  13791501
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.278e+05  9.678e+03    13.2   <2e-16 ***
## d$cost      2.362e-01  1.772e-03   133.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 685300 on 5519 degrees of freedom
## Multiple R-squared:  0.7631, Adjusted R-squared:  0.7631
## F-statistic: 1.778e+04 on 1 and 5519 DF,  p-value: < 2.2e-16
```

The least square regression line of *commitment* as x-axis and *cost* as y-axis is $y = 0.2362x + 127800$. The positive slope shows that the Ministry of Colleges and Universities is willing to invest more if a research project costs higher, as much as \$0.2362 for every \$1 research project cost. The $R^2 = 0.7631$ implies there is quite high correlation between *commitment* and *cost*, i.e. the correlation of investment by Ministry of Colleges and Universities and the actual research project cost. This can be interpreted as if a research project costs more, there is more investment towards it, but not as much. In other words, the more expensive a research project is, the proportion of the investment (*commitment*) towards the actual project cost (*cost*) increases but not as much.

5.2 Logistic Regression of *ratio* to *cost*

```
##
## Call:
## glm(formula = lr ~ t_lm$cost, family = binomial, data = t_lm)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.299e+00  1.986e-01   16.61   <2e-16 ***
## t_lm$cost   -1.363e-05  8.228e-07  -16.57   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2509.5  on 1992  degrees of freedom
## Residual deviance: 1335.2  on 1991  degrees of freedom
## AIC: 1339.2
##
## Number of Fisher Scoring iterations: 11
```

A logistic regression is done to analyze the previous interpretation, where a more expensive research project is likely to have lower *ratio* (i.e. lower proportion of investment, *commitment*). Setting the threshold of *ratio* at 0.5, the model produces $y = -0.00001363x + 3.299$ where the negative slope (-0.00001363) confirms the said interpretation, however its very small-sized magnitude shows a very weak correlation.

5.3 Cross-Validation

```
## [1] 0.038623
```

```
## [1] 0.0388748
```

Splitting the dataset into two parts (training dataset and testing dataset) with a proportion choice of 60%-40% respectively, the results are displayed as MSEs. As it appears, the training dataset and the testing dataset shows a very similar MSE, which is a positive sign.

6. Summary

This report analyzes how research project investments by Ministry of Colleges and Universities vary in Toronto. The analysis reveals that the Ministry of Colleges and Universities are willing to invest on research projects in general. Moreover, they are willing to invest more if a research project is more expensive, although the proportion of the investment is not as much as the rate of the increase in the cost of the research project. University of Toronto is found to be the institution with the most number of research projects in Toronto which made them the institution with the most *commitment* and *cost*. The spikes in *cost* (but with not as much *commitment*) shown in sub-section 3.2 is further explained by the variance testing in sub-section 4.5, where there is indeed less variance in *commitment* for the dominating institution (University of Toronto) and the opposite for other institutions. As a recommendation, the Ministry of Colleges and Universities might want to foster more research projects for other colleges and universities in Toronto as well as paying attention to their number of investment as an effort to make investments more even among institutions.

Appendix

Sub-Section 1.3

```
d = d %>% select(Lead.Research.Institution, Approval.Date, City, Ontario.Commitment,
                 Total.Project.Costs)
d = d %>% mutate(commitment = as.numeric(str_remove_all(str_remove_all(Ontario.Commitment, "\\$"), ",")
d = d %>% mutate(cost = as.numeric(str_remove_all(str_remove_all(Total.Project.Costs, "\\$"), ",")))
d = d %>% mutate(city = (str_remove_all(City, " ")))
d = d %>% mutate(institution = Lead.Research.Institution)
d = d %>% mutate(year = as.numeric(str_sub(Approval.Date, end = 4)))
d = d %>% mutate(ratio = commitment / cost)
dclean = d %>% select(Lead.Research.Institution, year, city, commitment, cost)
head(dclean, 8)
```

Sub-Section 2.1

```
DM = d %>% select(year, commitment, cost)
DM = DM %>% group_by(year) %>% summarise(commitment = sum(commitment), cost = sum(cost))
```

Sub-Section 2.2

```
CM = d %>% select(city, commitment, cost)
CM = CM %>% group_by(city) %>% summarise(commitment = sum(commitment), cost = sum(cost))
data2 = CM %>% arrange(desc(commitment))
```

Sub-Section 2.3

```
t = d %>% filter(city == "Toronto") %>% select(institution, commitment, cost)
t = t %>% filter(grepl("university", institution, ignore.case = TRUE) |
                   grepl("college", institution, ignore.case = TRUE))
t = t %>% filter(!grepl("hospital", institution, ignore.case = TRUE) &
                   !grepl("health", institution, ignore.case = TRUE))
t_box = t
t = t %>% group_by(institution) %>% summarise(N = n(), commitment = sum(commitment), cost = sum(cost))
t = t %>% arrange(desc(N))
data3 <- head(t, n = 8)
```

Sub-Section 2.4

```
t_avg = t %>% mutate(avg_commitment = commitment / N, avg_cost = cost / N) %>%
  mutate(avg_ratio = avg_commitment / avg_cost)
data4 = t_avg %>% select(institution, N, avg_commitment, avg_cost, avg_ratio)
```

Sub-Section 3.1

```
t_box = t_box %>%
  mutate(u = case_when(institution == "University of Toronto" ~ 1,
                       institution == "York University" ~ 2,
                       institution == "Ontario College of Art and Design (OCAD) University" ~ 4,
                       institution == "George Brown College of Applied Arts and Technology" ~ 5,
                       institution == "Royal Military College of Canada" ~ 6,
```

```
                        institution == "Humber College" ~ 7,
                        institution == "Seneca College of Applied Arts and Technology" ~ 8,
                        TRUE ~ 0)) %>%
  filter(u %in% c(1:8)) %>% mutate(ratio = commitment / cost)
data5 = ggplot(t_box, aes(group = u, x = u, y = ratio)) +
  geom_boxplot() + scale_x_continuous(breaks=1:8)
```

Sub-Section 3.2

```
t_curve = d %>% filter(city == "Toronto") %>%
  filter(institution %in% c("University of Toronto", "York University",
                            "Toronto Metropolitan University (formerly Ryerson University)",
                            "Ontario College of Art and Design (OCAD) University",
                            "George Brown College of Applied Arts and Technology",
                            "Royal Military College of Canada", "Humber College",
                            "Seneca College of Applied Arts and Technology")) %>%
  select(year, institution, commitment, cost)
t_curve_one = t_curve %>% group_by(year) %>% summarise(commitment = sum(commitment), cost = sum(cost))
t_curve_one <- pivot_longer(t_curve_one, cols = c(commitment, cost), names_to = "type",
                            values_to = "value")
data6 = ggplot(t_curve_one, aes(x = year, y = value, color = type)) + geom_line() +
  scale_y_continuous(labels = scales::dollar)
```

Sub-Section 3.3

```
t_curve_eight = t_curve %>% group_by(year, institution) %>%
  summarise(commitment = sum(commitment), cost = sum(cost), .groups = 'drop')
t_curve_eight <- pivot_longer(t_curve_eight, cols = c(commitment, cost),
                              names_to = "type", values_to = "value")
data7 = ggplot(t_curve_eight, aes(x = year, y = value, color = type)) +
  geom_line() +
  facet_wrap(~institution) +
  scale_y_continuous(labels = scales::dollar) +
  theme(axis.text.x = element_text(angle = 90))
```

Sub-Section 3.4

```
t_bar = d %>% filter(institution == "University of Toronto") %>%
  select(year, commitment, cost)
t_bar = d %>% group_by(year) %>%
  summarise(commitment = sum(commitment), cost = sum(cost))
t_bar <- pivot_longer(t_bar, cols = c(commitment, cost), names_to = "type",
                      values_to = "value")
data8 = ggplot(t_bar, aes(x = year, y = value, fill = type)) +
  geom_bar(position = "dodge", stat = "identity") +
  scale_y_continuous(labels = scales::dollar)
```

Sub-Section 4.1

```
t_t = d %>% filter(city == "Toronto")
t.test(t_t$ratio)
```

Sub-Section 4.2

```
boot_data = sample(t_t$ratio, size = nrow(t_t))
boot_function = function() {
  boot_s = sample(boot_data, size = nrow(t_t), replace = T)
  return(mean(boot_s))
  }
quantile(replicate(1000, boot_function()), c(0.025, 0.975))
```

Sub-Section 4.3

```
#prop-test
t_p = d %>% filter(city == "Toronto") %>% summarise(commitment = sum(commitment), cost = sum(cost))
prop.test(x = t_p$commitment, n = t_p$cost, p = 0.5)
```

Sub-Section 4.4

```
t_h = d %>% filter(city == "Toronto")
t.test(d$ratio, mu = 0.5, alternative = "greater", conf.level = 0.95)
```

Sub-Section 4.5

```
G1 = t_h %>% filter(institution == "University of Toronto")
G2 = t_h
response_var = c(G1$commitment, G2$commitment)
group_ind = c(rep("G1", length(G1$commitment)), rep("G2", length(G2$commitment)))
dd = data.frame(RV = response_var, group_ind = group_ind)
dd %>% group_by(group_ind) %>% summarize(group_mean = mean(RV), group_var = var(RV)) %>% as.data.frame
var.test(RV~group_ind, data = dd, alternative = "greater", conf.level = 0.95)
```

Sub-Section 5.1

```
model <- lm(d$commitment ~ d$cost, data = d)
summary(model)
```

Sub-Section 5.2

```
t_lm = d %>% filter(city == "Toronto")
t_lm = t_lm %>% mutate(lowratio = case_when(ratio < 0.5 ~ "insufficient",
                                            TRUE ~ "sufficient"))
t_lm = t_lm %>% mutate(lr = ifelse(lowratio=="sufficient",1,0))

model = glm(lr ~ t_lm$cost, family = binomial, data = t_lm)
summary(model)
```

Sub-Section 5.3

```
set.seed(123)
t_cv = d %>% filter(city == "Toronto")
t_cv = t_cv %>% mutate(group_ind = sample(c("train", "test"),
                                                    size=nrow(t_cv),
```

```
                                                    prob=c(0.6, 0.4),
                                                    replace = T))

#MSE for training dataset
m = lm(t_cv$ratio ~ t_cv$cost, data = t_cv %>% filter(group_ind=="train"))
y.hat=predict(m)

mean((t_cv$ratio[t_cv$group_ind=="train"] - y.hat)^2)

#MSE for testing dataset
y.hat=predict(m, newdata = t_cv %>% filter(group_ind=="test"))

mean((t_cv$ratio[t_cv$group_ind=="test"] - y.hat)^2)
```