STAA57 Final Project

Eishan Ashraf - 1010275499 Kshitij Kapoor - 1010297581 Daniel Venistan - 1010100506

April 04, 2025

Contents

1 Introduction

This is the beginning of our project report. We explore data from Ontario's Public Sector Salary Disclosure, commonly referred to as the "Sunshine List," from 1996 to 2020.

The Sunshine List is published annually by the Government of Ontario to disclose the names, positions, and salaries of certain public sector employees earning above a specific threshold, historically $100,000. Our goal is to investigate trends, compare salary distributions, and gain insights into how different public sector roles and organizations have evolved over time.

2 Merging and Cleaning Public Sector Salary Data (1996–2020)

To prepare our dataset for analysis, we gathered and merged Ontario's Public Sector Salary Disclosure data from multiple annual CSV files from the years **1996 to 2020**. These datasets changed slightly in structure across years, so we first ensured that all key fields—such as salary paid, taxable benefits, job title, and employer names were standardized and consistently named. Next, we cleaned the salary and benefits data by removing formatting symbols like dollar signs and commas and converting the entries into numerical values. We also refined the sector labels by correcting for inconsistencies such as extra characters, French-language suffixes, and formatting differences to ensure uniform naming conventions across years.

Only the relevant columns were retained: Sector, Last Name, First Name, Salary Paid, Taxable Benefits, Employer, Job Title, and Calendar Year. Each year's data was labeled appropriately and combined into one large dataset.

3 Description of the Dataset

We have annual CSV files from **1996** through **2020**, each containing salary disclosures for Ontario public sector employees. After cleaning, each dataset includes these key columns:

```
## [1] "Sector"           "Last.Name"        "First.Name"       "Salary.Paid"
## [5] "Taxable.Benefits" "Employer"         "Job.Title"        "Calendar.Year"
```

1. Sector: The broad classification of the public sector organization (e.g., Government, Hospitals, etc.).
2. Last.Name: The employee's last name.
3. First.Name: The employee's first name.
4. Salary.Paid: The total salary paid to the employee in that calendar year.
5. Taxable.Benefits/Benifits: The taxable benefits received by the employee.
6. Employer: The name of the employer or organization.
7. Job.Title: The employee's official position or title.
8. Calendar.Year/Year: The year for which the data applies (e.g., 1996, 1997, . . . , 2020).

4 Background of the Data

This dataset, covering the years 1996 to 2020, was collected under the Public Sector Salary Disclosure Act, which requires the public release of compensation details for employees earning $100,000 or more. The data is published annually by the Government of Ontario and includes comprehensive information on **each employee's salary, taxable benefits, job title, employer, and sector**.

Through this long-term record, researchers, policymakers, journalists, and the public can explore trends in public sector compensation, transparency, and workforce dynamics. By spanning multiple decades, the dataset provides insights into how salaries, benefits, and job roles have evolved over time in Ontario's public sector.

Ultimately, this dataset offers a valuable resource for understanding long-term compensation trends, informing policy decisions, and enhancing transparency in government spending. It can be used to investigate salary changes or assess the competitiveness of public sector roles.

5 What is the overall Research Question

As researchers, our overarching goal is to explore the patterns and trends in public sector compensation in Ontario from **1996 to 2020**. We aim to analyze various factors, including salary, taxable benefits, sector, and year, to gain insights into how salary distributions have evolved across different public sector organizations. By examining these data, we hope to identify meaningful patterns, compare compensation, and potentially forecast future salary trends. Specifically, our research questions include:

1. **What are the top-paying sectors by year?**

2. **Do colleges and universities pay similar average salaries, or is there a notable difference?**

3. **How have salary trends changed by sector over the years?**

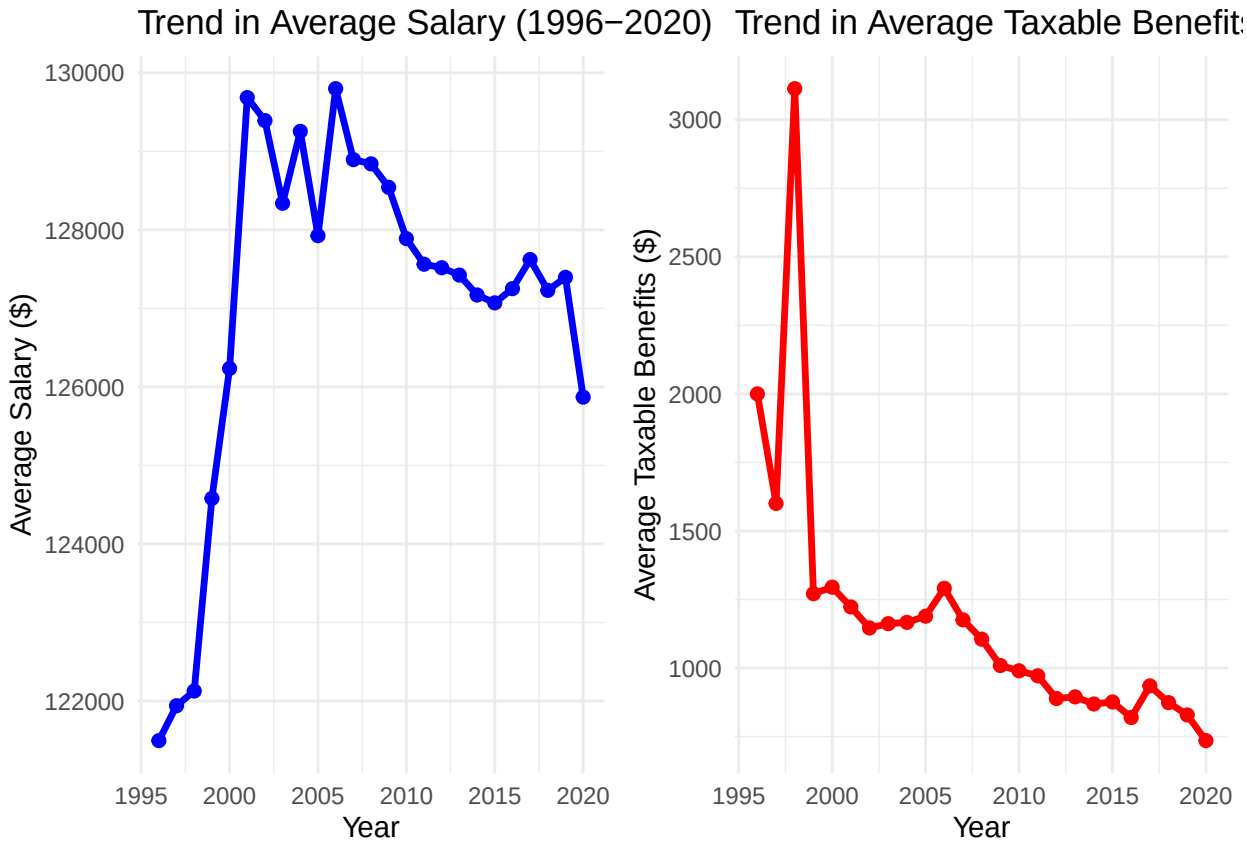
4. **What are the average salary and taxable benefits for each year from 1996 to 2020?**

5. **Can we predict the average salary for upcoming years based on historical data?**

By addressing these questions, we aim to provide a clearer understanding of how public sector compensation has changed over time and what factors may influence these patterns.

6 Tables and Graphs

6.1 Average Salary and Taxable Benefits by Year (1996-2020)



The first graph, "Trend in Average Salary (1996-2020)", displays how the average salary of public sector employees on the Sunshine List has evolved over time. In this plot, the x-axis represents the calendar year while the y-axis shows the average salary in dollars. Notably, despite some minor annual fluctuations, **the average salary has remained consistent, generally ranging between $120,000 and $130,000**. For instance, the average salary was approximately $121,495 in 1996 and increased slightly to around $125,870 in 2020. This consistency suggests that, **even in the face of inflation, economic downturns (such as the 2008 recession), and shifts in government policies, the base salaries for high-earning public sector employees have essentially remained unchanged**.

In contrast, the second graph, "Trend in Average Taxable Benefits (1996-2020)", illustrates a **clear downward trend in the average taxable benefits over the same period**. Here, the x-axis again represents the year, and the y-axis represents the average taxable benefits in dollars. While the average taxable benefits were around $2,000 in 1996, they declined steadily to approximately $735 by 2020. This reduction implies that although base salaries have been generally the same and stable, the additional monetary perks provided as taxable benefits have decreased over time.

Overall, these findings highlight a significant change in the total compensation for high-earning public sector employees. While the **base salaries have been stable, there is a notable reduction in taxable benefits** suggesting that the **overall total compensation** when combining salary and taxable benefits has **decreased over time**. One potential explanation for this trend is that public sector organizations may be actively trying to decrease costs by reducing supplementary benefits.

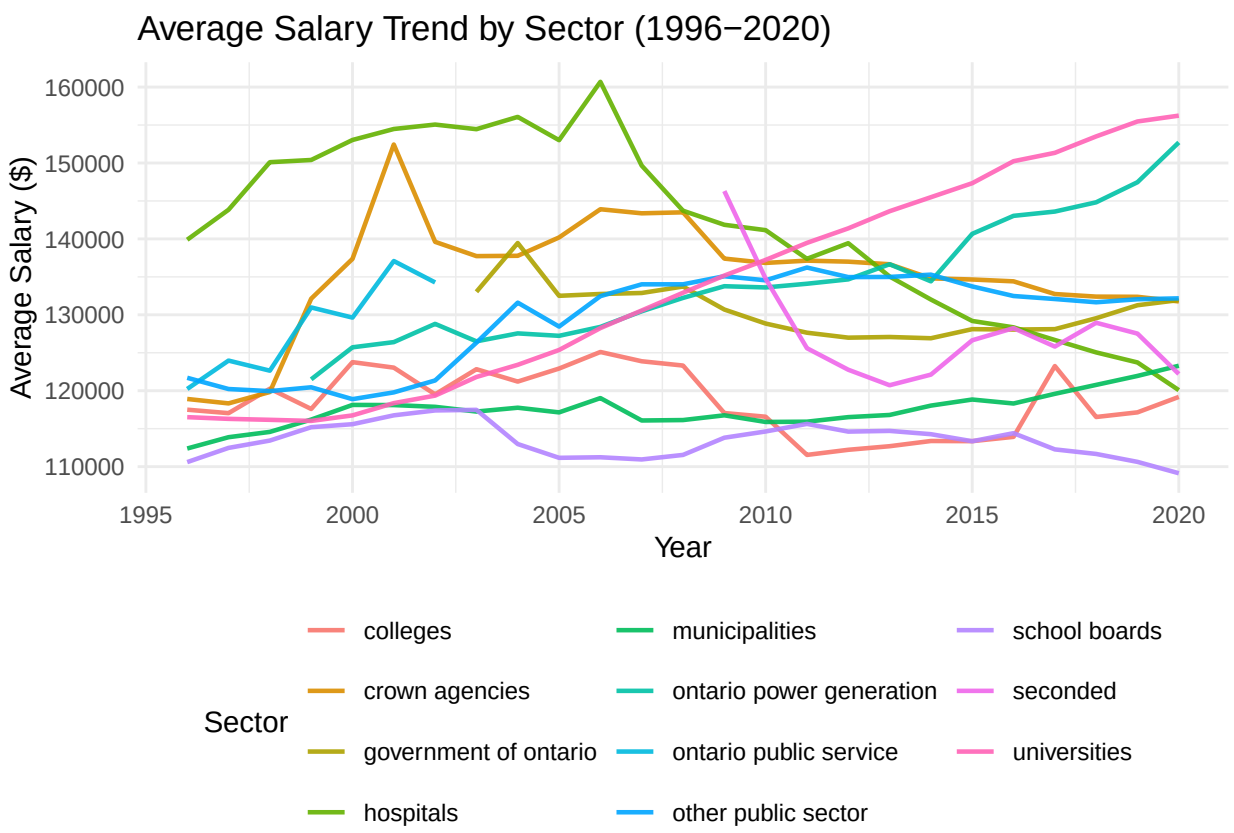
6.2 Sectors with the Largest Average Salary by Year

Table 1: Sector with the Largest Average Salary by Year (1996-2020)

Sector	Calendar.Year	Avg.Salary
hospitals	1996	139890.5
hospitals	1997	143842.0
hospitals	1998	150105.3
hospitals	1999	150397.0
hospitals	2000	153031.1
hospitals	2001	154476.4
hospitals	2002	155059.3
government of ontario - judiciary	2003	184822.6
government of ontario - judiciary	2004	244582.8
government of ontario - judiciary	2005	164164.9
government of ontario - judiciary	2006	207092.0
government of ontario - judiciary	2007	215279.3
government of ontario - judiciary	2008	199401.9
seconded (environment)	2009	313967.3
seconded (health and long-term care)	2010	326929.8
seconded (economic development and innovation)	2011	263450.2
seconded (health and long-term care)	2012	346040.9
government of ontario - judiciary	2013	190294.5
seconded (health and long-term care)	2014	302052.3
seconded (health and long-term care)	2015	224621.2
government of ontario - judiciary	2016	207014.6
government of ontario - judiciary	2017	204664.9
government of ontario - judiciary	2018	210775.7
government of ontario - judiciary	2019	217121.0
government of ontario - judiciary	2020	234147.2

The table titled “Sector with the Largest Average Salary by Year (1996–2020)” provides a clear summary of which public sector categories offered the highest average annual salaries in Ontario from 1996 through 2020. The column “Sector” identifies the specific public sector category or department, such as Hospitals, the Judiciary, or specialized Seconded positions within government ministries. The “Calendar Year” column indicates the respective year in which the average salaries were recorded, while the “Avg. Salary” column lists the corresponding highest average salaries in Canadian dollars. However, the sector with the highest average salary changed over time. Initially, between 1996 and 2002, hospitals consistently had the highes average salary. Beginning in 2003, however, the Judiciary sector became prominent. Between 2009 and 2015, certain specialized “Seconded” positions—temporary roles transferred to key ministries, mostly Health & Long-Term Care—displayed exceptionally high average salaries, peaking at approximately $346,041 in 2012. Post 2015, the Judiciary sector regained its position as the top-paying sector, consistently maintaining high salary averages through 2020.

6.3 Salary trends by sector over the years



This is a graph that shows the average salary of the sectors throughout the years. The sectors (written on the y-axis) are the dependent variables which rely on the independent variable (written on the x-axis), the year from which the data was sampled. The **data was simplified to 11 sectors (initially having 50+ different sectors) by combining sectors that had similar governing bodies or were similar in nature**. For example, the sectors "hydro one and ontario power generation" and "ontario power generation" are very similar, so combining them and averaging out their salaries would make for a cleaner and more readable graph. The same process was used for Hospital-related jobs, the specialized Seconded jobs, other Public Sector jobs, and Municipality-related jobs. This grouping of data makes it easier to see the trends of the salaries for every sector, allowing us to gain valuable insight.

Over the years, there has been a change of the sectors with highest average salaries. Most notably, **both the University and the Ontario Power Government sectors have had average salaries that started the 25-year period near the bottom of the rankings and ended it as the top 2 highest paying sectors**. Both sectors had a significant increase in their average salaries. Furthermore, the **Hospital-related jobs had, on average, the highest paying jobs in 1996 by a large margin**. The salary of these jobs also only rose in the next few years; however, this is where it began declining. **Post 2006 and onwards, the Hospital sector began rapidly declining** and the jobs were not paying as much as they once were. By the year 2020, the average salary of jobs in the Hospital sector had fallen to close to the bottom of all sectors shown here.

Looking at the graph, we can see that, on average, **universities have always paid their staff more colleges have**. Around the late 90's, we can see that the average salary of the university staff was less than the average salary of the college staff, but that did not last long. Ever since then, and as society has evolved over the past 2 decades, we have seen **rapid and consistent growth of the average salary of university workers**. In recent years, **it has also overtaken all other sectors as the highest paying sector**. **The salary for college workers has stayed relatively consistent through the years, causing it to fall well behind the average salary that universities can offer**.

We can note that there is **no consistent salary trend that is seen over the years for every sector**. Every sector has had its own trend. The **average pay for some sectors has fallen, others have had their pay increase, while lastly, there is a group of sectors which have had their pay stay consistent**, and we have spoken shown example of a sector falling in each type. As a novice, one may believe that with the rising inflation, current downfall of the Canadian currency, and the increasing cost to live, the salaries of every Public Sector job should also be increasing at some steady rate; however, we see this is not the case. These factors that should be ground for Public Sector jobs getting consistent raises to

their salaries pose many threats to current Canadians, and the inconsistent overall trends that we see in the graph above do not paint a picture for a positive future.

7 Hypothesis Testing

7.1 Average salary in Colleges vs Average salaries in University from 1996 to 2020.

Table 2: Average Salaries for Colleges and Universities (1996–2020) and Their Differences(College - Universities)

Calendar.Year	colleges	universities	Diff
1996	117502.0	116517.3	984.7340
1997	117045.4	116295.0	750.3855
1998	120264.4	116168.4	4096.0538
1999	117595.5	116022.3	1573.1693
2000	123766.5	116750.3	7016.1686
2001	123043.1	118367.1	4675.9591
2002	119521.4	119356.4	164.9405
2003	122833.4	121809.4	1024.0444
2004	121202.9	123414.8	-2211.9091
2005	122921.6	125367.9	-2446.3075
2006	125104.5	128250.8	-3146.2753
2007	123881.5	130568.3	-6686.8498
2008	123313.9	132906.3	-9592.4088
2009	117059.2	135171.6	-18112.3747
2010	116574.7	137233.0	-20658.2779
2011	111543.8	139472.6	-27928.8141
2012	112212.5	141389.9	-29177.4406
2013	112688.8	143642.3	-30953.4709
2014	113375.8	145492.9	-32117.0869
2015	113357.4	147334.3	-33976.8853
2016	113921.0	150235.6	-36314.6472
2017	123248.3	151336.3	-28088.0601
2018	116539.6	153502.1	-36962.4796
2019	117177.0	155469.2	-38292.2122
2020	119197.9	156238.1	-37040.2505

We want to determine whether colleges pay more on average than universities in Ontario, when considering data from 1996 through 2020. Instead of looking at all salaries across multiple years which can be influenced by inflation or other time related factors we **compute the average salary for colleges and universities for each year from 1996 to 2020 and take their yearly difference**, d_i = (average college for year i) − (average university salary for year i) where i indicates the current year. So we have that d_i denotes the average salary difference between colleges and university for each year i we can define the following.

Hypothesis:

- **Null Hypothesis (H_0)**: On average, there is **no difference** in salaries between college and university employees from 1996 to 2020. That is, the true mean of all yearly differences is zero:

$$H_0 : \mu_d = 0$$

- **Alternative Hypothesis (H_1)**: On average, there is a difference in salaries between college and university employees from 1996 to 2020. That is, the true mean of all yearly differences is not zero:

$$H_1 : \mu_d \neq 0$$

To test this hypothesis, we can use a two-sample t-test to determine whether colleges pay more on average than universities in Ontario, using the Diff column (which represents the yearly salary differences) from 1996 through 2020. We choose a

significance level of $\alpha = 0.05$. The p-value is the probability of observing a result as extreme or more extreme than what we found, assuming the null hypothesis H_0 is true. If the p-value is less than our chosen α, we reject the null hypothesis. We can also use confidence intervals to assess the difference. A 95% confidence interval for the mean difference tells us the range of values within which we are 95% confident the true mean difference lies. If the confidence interval includes 0, this suggests there may be no significant difference, supporting the null hypothesis. If the confidence interval does not include 0, this supports the alternative hypothesis suggesting a possible significant difference in average salaries.

```
##
##  One Sample t-test
##
## data:  avg_salaries$Diff
## t = -4.5328, df = 24, p-value = 0.0001362
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -21737.849  -8135.774
## sample estimates:
## mean of x
## -14936.81
```

The test showed a ***t*-statistic of -4.5328 with 24 degrees of freedom and a very small *p*-value of 0.0001362**. Since the p-value is far below our chosen significance level of 0.05, **we reject the null hypothesis**. In addition, the 95% confidence interval for the mean difference is [-21737.849, -8135.774], which does not contain zero. This further supports the that the **average salary difference over these years is significantly different from zero**. The negative sample mean of the differences $\bar{d} = -14,936.81$ suggests that, **on average, universities paid more than colleges from 1996 to 2020**. Since the p-value is very small and the 95% confidence interval for the mean difference does not include zero, we find strong evidence of a real difference in average salaries between the two sectors.

8 Boot Strapping

We have used bootstrapping to estimate the average salary, and tax benefits from 1996 to 2020. We also calculated a 95% confidence interval. To do so we took a sample of 50 from the dataset, then replicated the a sample with repeats 100000 times. Each time we took the mean of the sample. Using this we calculated the mean and confidence interval for the data. This was performed twice, once for the salary and once for the tax benefits.

```
## The average salary paid from 1996 to 2020: 120222.3
```

```
## 95% confidence interval for salary: ( 115162.4 - 125701.5 )
```

```
## The average tax benefit from 1996 to 2020: 1475.2
```

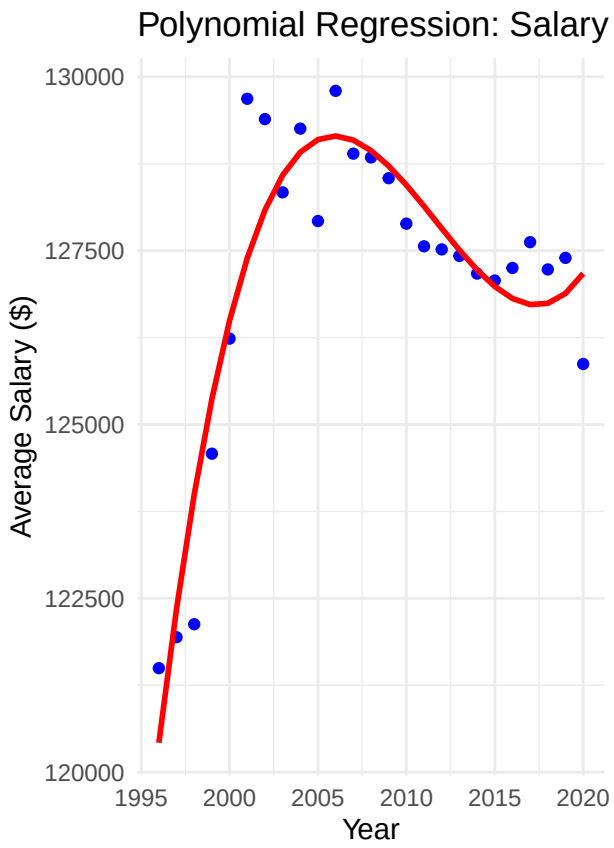
```
## 95% confidence interval for tax benefit: ( 423.0186 - 3368.47 )
```

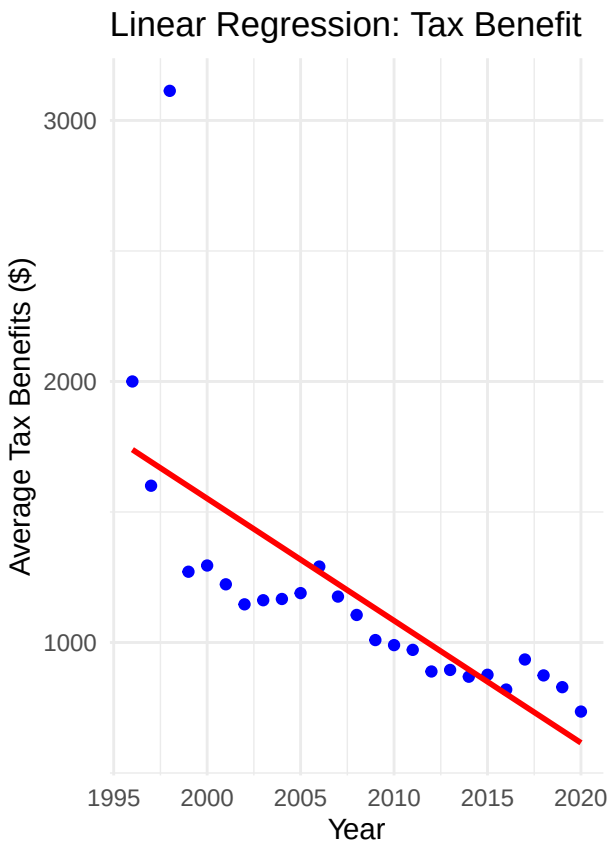
This shows that the **average salary paid from 1996 to 2020 was $120222.30**. Since the 95% confidence interval for salary is ($115162.4 - $125701.5), we are 95% sure that the actual mean is within that range. Similarly, \textbf{the average tax benefit from 1996 to 2020 was $1475.20}.

9 Machine Learning Model

We wanted to create 2 models, one to predict the average salary over years, and predict tax benefits over years. We will be treating the years as a continuous independent variable, which means the tax benefit and the salary will be the dependent variables for their respective models.

9.1 Model Graphs






```
##
## Call:
## lm(formula = Avg_Salary ~ poly(Calendar.Year, 3), data = avg_table)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1866.29  -411.23   -98.34   487.21  2290.95
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              127082.2      187.4 678.096  < 2e-16 ***
## poly(Calendar.Year, 3)1   4178.0      937.1   4.459 0.000217 ***
## poly(Calendar.Year, 3)2  -8285.4      937.1  -8.842 1.60e-08 ***
## poly(Calendar.Year, 3)3   4771.1      937.1   5.092 4.83e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 937.1 on 21 degrees of freedom
## Multiple R-squared:  0.8552, Adjusted R-squared:  0.8345
## F-statistic: 41.33 on 3 and 21 DF,  p-value: 5.453e-09
```

9.2 Model Summaries

We have made a non-linear model with degree 3 to predict the salary. \textbf{Since the Multiple R squared value is 0.85, this means that the model explains 85% of the variance in the data}. **Since all p-values for the coefficients are very small, they are all significant so all coefficients are needed**.

```
##
```

```
## Call:
## lm(formula = Avg_Tax_Benefits ~ Calendar.Year, data = avg_table)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -327.36 -128.55  -65.05   26.68 1468.13
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   95161.751  19607.072   4.853 6.71e-05 ***
## Calendar.Year   -46.805      9.764  -4.793 7.79e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 352.1 on 23 degrees of freedom
## Multiple R-squared:  0.4997, Adjusted R-squared:  0.478
## F-statistic: 22.98 on 1 and 23 DF,  p-value: 7.791e-05
```

We have made a linear model to predict the tax benefits. \textbf{Since the Multiple R squared value is 0.499, this means that the model explains 50%% of the variance in the data}. **Since the p-values for the coefficients are really small thay are all significant. This means that all of the coefficients are needed**.

9.3 Interpretation of Regression Parameters

For the **first regression model**, we have the coefficients:

- The **intercept coefficient, 127082.2**, which tells us that if all independent variables are equal to 0, the **expected value for salary should be approximately 127082.2**.
- The coefficient for the **linear term is 4178.0**. This means, if we let the cubic and quadratic terms be constant, **the salary should increase by $4178.00 each year**.
- The **second degree coefficient is -8285.4**. This means that we have **a much larger dip in the equation, so the local minimum is much smaller**.
- The **third degree coefficient is 4771.1**. This means there our model is generally **concave up**.

For the **second regression model**, we have the coefficients:

- The **intercept coefficient, 95161.75**, which tells is that if all independent variables are zero, the **expected value for the tax benefits is approximately $95161.75**.
- The **coefficient for the linear term is -46.805**. This means the tax benefit should \textbf{decrease by $46.805}, every year.

10 Cross Validation

We used 5 fold cross validation for our models to train and test them. To do so we split the data into 5 parts, and the model was trained and tested 5 times. Each time a different group of data was used to test, and the other data was used to train. We then **used the root mean squared error (RMSE) method** to determine the accuracy of the model. Each time the RMSE was calculated and stored, which was finally averages to find the real accuracy of the models.

```
## The salary prediction model was on average off by $ 1081.36
```

```
## The tax prediction model was on average off by $ 305.62
```

Considering that the salaries and tax benefits are **quite large numbers**, having our model be off by such small values relatively **shows that our model is fairly accurate**.

11 Final Summary

Based on our findings in this report, **we can conclude the following**:

- On average, the sectors with the largest paying salaries have been the Hospital, University, and Ontario Power Government sectors. In the beginning of the time frame (closer to 1996), the Hospital sector topped the charts for average salary by a large margin, but as the years progressed, it has fallen quite drastically. Concurrently, the University and Ontario Power Government sectors have risen to the top. The other sectors have remained quite similar over the 25 years.
- Universities have, on average, paid their workers more than colleges have.
- There is no common trend amongst the sectors. Generally, the salary for each sector rises, falls, or stays consistent independently of the other sectors.
- The average taxable benefits of Public Sector workers have decreased over the years, while their average salaries have not increased enough to make up the difference. This trend looks likely to continue into future years, which can cause massive impacts on the workers.
- The salary prediction model was very accurate, with the expected values being approximately $1000 off of their respective real values. Similarly, the tax benefit model's expected values only deviated from the real values by approximately $300. These deviations may seem large and inaccurate at first, but since we are working with very large salaries, these deviations are very minuscule. Thus, we can predict future salaries and taxable benefits in the Public Government Sector fields with high precision.

Overall, we gained insight on **average salaries and taxable benefits per sector per year**. We also learned about any **trends respective to the sectors**, ensuring we can **predict future values for salaries and taxable benefits**.

12 Appendix

```
#1. Introduction: No code for this Part

#2.1 Merging and Cleaning Public Sector Salary Data (1996-2020) loading libraries
#Loading Ontario's Public Sector Salary Disclosure data for all the years
library(tidyverse)
library(dplyr)
library(knitr)
library(ggplot2)
library(gridExtra)

set2020 =  read.csv("en-2020-pssd-compendium.csv")
set2019 =  read.csv("en-2019-pssd-compendium (1).csv")
set2018 =  read.csv("en-2018-pssd-compendium.csv")
set2017 =  read.csv("en-2017-pssd-compendium.csv")
set2016 =  read.csv("en-2016-pssd-compendium.csv")
set2015 =  read.csv("en-2015-pssd-compendium-with-addendum (1).csv")
set2014 =  read.csv("en-2014-pssd-full-compendium.csv")
set2013 =  read.csv("en-2013-pssd.csv")
set2012 =  read.csv("en-2012-pssd.csv")
set2011 =  read.csv("en-2011-pssd.csv")
set2010 =  read.csv("en-2010-pssd.csv")
set2009 =  read.csv("en-2009-pssd.csv")
set2008 =  read.csv("en-2008-pssd.csv")
set2007 =  read.csv("en-2007-pssd.csv")
set2006 =  read.csv("en-2006-pssd.csv")
set2005 =  read.csv("en-2005-pssd.csv")
set2004 =  read.csv("en-2004-pssd.csv")
set2003 =  read.csv("en-2003-pssd.csv")
```

```r
set2002 =  read.csv("en-2002-pssd.csv")
set2001 =  read.csv("2001.csv")
set2000 =  read.csv("en-2000-pssd.csv")
set1999 =  read.csv("1999.csv")
set1998 =  read.csv("en-1998-pssd.csv")
set1997 =  read.csv("en-1997-pssd.csv")
set1996 =  read.csv("en-1996-pssd.csv")

#2.2 This function standardizes and filters salary dataset columns for consistency and analysis
filter_data <- function(dataset) {
  if (is.character(dataset$Salary.Paid)) {
    # Removes commas and $ from string
    dataset$Salary.Paid <- as.numeric(gsub("[$,]", "", dataset$Salary.Paid))
  }
  if (is.character(dataset$Taxable.Benefits)) {
    dataset$Taxable.Benefits <- as.numeric(gsub("[$,]", "", dataset$Taxable.Benefits))
  }
  dataset$Sector <- tolower(dataset$Sector) # To ensure for consistent Sector names
  # Replace all instances of one substring with the other
  # to ensure consistent grouping later on
  dataset$Sector <- gsub("[&]", "and", dataset$Sector)
  dataset$Sector <- gsub("[:]", "-", dataset$Sector)
  dataset$Sector <- gsub(" - universités", "", dataset$Sector)
  dataset$Sector <- gsub("[*]", "", dataset$Sector)
  dataset$Sector <- gsub("[\u2010-\u2015]", "-", dataset$Sector) # Standardize all dashes
  dataset$Sector <- enc2utf8(dataset$Sector) # Use UTF-8 Formatting
  dataset$Sector <- str_trim(dataset$Sector) # Remove trailing and leading white space
  return(dataset %>% select(Sector, Last.Name, First.Name, Salary.Paid,
                            Taxable.Benefits, Employer, Job.Title, Calendar.Year))
}

#2.3 For each year we changed column names and cleaned the data by removing formatting from
# numeric values, standardizing sector names, and selecting only the relevant columns:
# We define the final columns names to be Sector, Last.Name, First.Name, Salary.Paid,
# Taxable.Benefits, Employer, Job.Title, and Calendar.Year.
set2020 <- set2020 %>% rename(Last.Name = Last.name,First.Name = First.name,
                              Salary.Paid = Salary, Taxable.Benefits = Benefits,
                              Job.Title = Job.title, Calendar.Year = Year)

set2014 <- set2014 %>% rename(Last.Name = Last.name,Job.Title = Job.title,
                              Calendar.Year = Calendar.year)
set2001 <- set2001 %>% rename(Last.Name = Surname, Job.Title = Position)

# Need to do this so ensure Calendar.Year is numeric and binding works
set1999$Calendar.Year <- as.double(set1999$Calendar.Year)
# Placing all datasets into a named list with year as the name
data_list <- list(
  "2020" = set2020,"2019" = set2019,"2018" = set2018, "2017" = set2017,
  "2016" = set2016,"2015" = set2015,"2014" = set2014,"2013" = set2013,
  "2012" = set2012,"2011" = set2011,"2010" = set2010,
  "2009" = set2009,"2008" = set2008,"2007" = set2007,"2006" = set2006,
  "2005" = set2005, "2004" = set2004, "2003" = set2003, "2002" = set2002,
  "2001" = set2001,"2000" = set2000,"1999" = set1999,"1998" = set1998,
  "1997" = set1997,"1996" = set1996
)

# Iterate each years data set and clean it
for (year_str in names(data_list)) {
  # Convert list name to numeric year (e.g. "2019" -> 2019)
```

```r
  numeric_year <- as.numeric(year_str)
  # Extract the data frame
  curr_data <- data_list[[year_str]]
  # Assign correct Calendar.Year as some are missing
  curr_data$Calendar.Year <- numeric_year
  # Apply filter_data() function to df
  curr_data <- filter_data(df)
  # Store cleaned and filtered df back into list
  data_list[[year_str]] <- curr_data
}

# Directly bind everything in the list
master_set <- bind_rows(data_list)
# master_set now contains all rows from 1996 to 2020

#3. Description of the Dataset
names(set2020)
names(set2019)

#4. Background of the Data: No Code in this Part

#5. What is the overall Research Question: No code for this Part

#6. Tables and Graphs
#6.1 Average Salary and Taxable Benefits by Year (1996-2020)
# Group by Calendar.Year and compute the averages
avg_table = master_set %>%
  group_by(Calendar.Year) %>%
  summarize(
    Avg_Salary = mean(Salary.Paid, na.rm = TRUE),
    Avg_Tax_Benefits = mean(Taxable.Benefits, na.rm = TRUE),
    .groups = "drop"
  ) %>%
  arrange(Calendar.Year)

# Print the table
kable(avg_table,
      caption = "Summary of Average Salary and Taxable Benefits by Year (1996-2020)",
      digits = 2)

# Create Average Salary plot
plot1 <- ggplot(avg_table, aes(x = Calendar.Year, y = Avg_Salary)) +
  geom_line(color = "blue", linewidth = 1.2) +
  geom_point(color = "blue", size = 2) +
  labs(
    title = "Trend in Average Salary (1996-2020)",
    x = "Year",
    y = "Average Salary ($)"
  ) +
  theme_minimal()

# Create the Average Taxable Benefits plot
plot2 <- ggplot(avg_table, aes(x = Calendar.Year, y = Avg_Tax_Benefits)) +
  geom_line(color = "red", linewidth = 1.2) +
  geom_point(color = "red", size = 2) +
  labs(
    title = "Trend in Average Taxable Benefits (1996-2020)",
    x = "Year",
    y = "Average Taxable Benefits ($)"
```

```
  ) +
  theme_minimal()

# Arrange the two plots side by side
grid.arrange(plot1, plot2, ncol = 2)

#6.2 Sectors with the Largest Average Salary by Year
# Grouping data by Sector and Year then calculating average salary per group and sort in
# ascending order
master_set2 = master_set %>% group_by(Sector, Calendar.Year) %>%
  summarize(Avg.Salary = mean(Salary.Paid, na.rm = TRUE)) %>%
  group_by(Calendar.Year) %>% filter(Avg.Salary == max(Avg.Salary, na.rm = TRUE)) %>%
  arrange(Calendar.Year)

# Displaying final table
kable(master_set2,
      caption = "Sector with the Largest Average Salary by Year (1996-2020)",
      booktabs = TRUE,
      digits = 2)

#6.3 Salary trends by sector over the years
# Copy master_set
per_sector_set = master_set

# Rename sectors to ensure correct grouping
per_sector_set$Sector <- per_sector_set$Sector %>%
  str_to_lower() %>%
  str_trim() %>%
  str_replace("^government of ontario.*", "government of ontario") %>%
  str_replace("^ministry.*", "ministry") %>%
  str_replace("^municipalities.*", "municipalities") %>%
  str_replace("^seconded.*", "seconded") %>%
  str_replace("^hospitals.*", "hospitals") %>%
  str_replace("hydro one and ontario power generation", "ontario power generation") %>%
  str_replace("other public sector employers", "other public sector")

# Group by Sector and Calendar.Year and take average of Salary.Paid
per_sector_set <- per_sector_set %>%
  group_by(Sector, Calendar.Year) %>%
  summarise(Avg_Salary = mean(Salary.Paid, na.rm = TRUE))

# Plot graph
ggplot(per_sector_set, aes(x = Calendar.Year, y = Avg_Salary, color = Sector)) +
  geom_line(size = 0.8, alpha = 0.9) +
  xlab("Year") + ylab("Average Salary") +
  ggtitle("Average Salary Trend by Sector (1996-2020)") + labs(
    color = "Sector") +
  theme_minimal() +
  theme(
    legend.position = "bottom") +
  guides(color = guide_legend(ncol = 3))

#7. Hypothesis Testing
# Filter & standardize sector names that start or contain college and universit
college_uni_data = master_set %>%
  filter(grepl("college", Sector, ignore.case = TRUE) |
         grepl("universit", Sector, ignore.case = TRUE)) %>%
  mutate(Sector = case_when(
    grepl("college", Sector, ignore.case = TRUE) ~ "colleges",
```

```
    grepl("universit", Sector, ignore.case = TRUE) ~ "universities",
    TRUE ~ Sector))

# Compute average salary by sector and year
avg_salaries = college_uni_data %>%
  group_by(Calendar.Year, Sector) %>%
  summarise(Avg_Salary = mean(Salary.Paid, na.rm = TRUE), .groups = "drop") %>%
  pivot_wider(names_from = Sector, values_from = Avg_Salary)

# Compute difference for each year: colleges - universities and make it a new column
avg_salaries = avg_salaries %>%
  mutate(Diff = colleges - universities)

#7.1 Average Salaries for Colleges and Universities (1996-2020) and
# their Differences(College - Universities)
kable(avg_salaries, caption = "Average Salaries for Colleges and Universities (1996-2020) and
      their Differences(College - Universities)")

#8. Boot Strapping
set.seed(999)
obs.sam.sal=sample(master_set$Salary.Paid,size=50) # Observed sample
obs.sam.tax=sample(master_set$Taxable.Benefit, size=50)

boot_function=function(d){
  boot_s = sample(d, size=50, replace=TRUE)
  return(mean(boot_s))
}
boot_sal = replicate(100000,boot_function(obs.sam.sal)) # Fake replication of samples
boot_tax = replicate(100000,boot_function(obs.sam.tax))
#95%
conf.sal = quantile(boot_sal, c(0.025,0.975))
conf.tax = quantile(boot_tax, c(0.025,0.975))

cat("The average salary paid from 1996 to 2020:", mean(boot_sal), "\n")
cat("95% confidence interval for salary:", "(", conf.sal[1],"-", conf.sal[2],")", "\n\n")
cat("The average tax benefit from 1996 to 2020:", mean(boot_tax), "\n")
cat("95% confidence interval for tax benefit:", "(", conf.tax[1],"-", conf.tax[2],")", "\n")

#9. Machine Learning Model

##9.1 Model Graphs
# Fit a quadratic (2nd-degree) polynomial regression model
model <- lm(Avg_Salary ~ poly(Calendar.Year, 3), data=avg_table)

# Add predicted values to the table
avg_table$predicted_salary <- predict(model, newdata=avg_table)

# Plot the results
model1 = ggplot(avg_table, aes(x=Calendar.Year, y=Avg_Salary)) +
  geom_point(color="blue") +  # Scatter plot of actual salary data
  geom_line(aes(y=predicted_salary), color="red", size=1) +  # Regression line
  ggtitle("Polynomial Regression: Salary Trend") +
  xlab("Year") + ylab("Average Salary ($)") +
  theme_minimal()

# Fit a quadratic (2nd-degree) polynomial regression model
model_tax <- lm(Avg_Tax_Benefits ~ Calendar.Year, data=avg_table)

# Add predicted values to the table
```

```r
avg_table$predicted_tax <- predict(model_tax, newdata=avg_table)

# Plot the results
model2 = ggplot(avg_table, aes(x=Calendar.Year, y=Avg_Tax_Benefits)) +
  geom_point(color="blue") +  # Scatter plot of actual salary data
  geom_line(aes(y=predicted_tax), color="red", size=1) +  # Regression line
  ggtitle("Polynomial Regression: Salary Trend") +
  xlab("Year") + ylab("Average Salary ($)") +
  theme_minimal()

grid.arrange(model1, model2, ncol = 2)

summary(model)

##9.2 Model Summaries
summary(model_tax)

##9.3 Interpretation of Regression Parameters: No code for this Part

#10. Cross Validation:
# Using 5 fold validation
k=5

# Preparing k samples
avg_table_cv <- avg_table %>% select(Calendar.Year, Avg_Salary, Avg_Tax_Benefits)
avg_table_k = avg_table_cv %>% mutate(idx = sample(c(1:k), size=nrow(avg_table_cv), replace = TRUE))

avg_table_k = avg_table %>% mutate(idx = sample(c(1:k), size=nrow(avg_table), replace = T))

sal_index = vector()
tax_index = vector()

# Training and testing on each of the samples
for (i in 1:k){
  # Preparing train and test data
  data_train = avg_table_k %>% filter(idx != i)
  data_test = avg_table_k %>% filter(idx == i)

  # Validating salary model
  sal_trained_model = lm(Avg_Salary ~ poly(Calendar.Year, 3), data=data_train)
  sal_test_result = predict(sal_trained_model, newdata=data_test)
  sal_mse <- mean((sal_test_result - data_test$Avg_Salary)^2)
  sal_index[i]=sqrt(sal_mse)

  # Validating tax model
  tax_trained_model = lm(Avg_Tax_Benefits ~ poly(Calendar.Year, 3), data=data_train)
  tax_test_result = predict(tax_trained_model, newdata=data_test)
  tax_mse <- mean((tax_test_result - data_test$Avg_Tax_Benefits)^2)
  tax_index[i]=sqrt(tax_mse)
}

# Output final RMSE
cat("The salary prediction model was on average off by $",mean(sal_index), "\n")
cat("The tax prediction model was on average off by $", mean(tax_index))

#11. Final Summary: No code for this Part

#12. Summary: Everything in this section!
```