# Individual Assignment 1: Data Exploration and Visualization by Shimony Agrawal

Download the necessary packages for data exploration and visualization.

```
# install.packages("DBI")
# install.packages("odbc")
# install.packages("dplyr")
# install.packages("tidyverse")
# install.packages("lubridate")
# install.packages("ggplot2")

library(DBI)
```

```
## Warning: package 'DBI' was built under R version 4.0.2
```

```
library(odbc)
```

```
## Warning: package 'odbc' was built under R version 4.0.2
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.2
```

```
## -- Attaching packages ---------------------------------------------------------------------- tidyverse
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.1     v dplyr   1.0.0
## v tidyr   1.1.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## Warning: package 'ggplot2' was built under R version 4.0.2
```

```
## Warning: package 'dplyr' was built under R version 4.0.2
```

```
## Warning: package 'forcats' was built under R version 4.0.2
```

```
## -- Conflicts ------------------------------------------------------------------------ tidyverse_conflicts
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

```
## Warning: package 'lubridate' was built under R version 4.0.2
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(dplyr)
library(ggplot2)
```

Task 1: Download the file 'nycpay.csv' from our class Blackboard site. This dataset contains salary information for every New York City municipal employee for the fiscal years 2014-2019.

```
NYC_Pay = read.csv("/Users/shimonyagrawal/Desktop/Grad /Summer 2/AD699_Data Mining/RStudio/Assignment 1/
```

Task 2: How many rows and how many columns does your dataframe contain?

```
dim(NYC_Pay)
```

```
## [1] 3333080      17
```

Comments: The dataframe has 3,333,080 rows and 17 columns.

Task 3: Filter the dataframe. Create a new object that only contains data for your assigned year and borough. How many rows does your dataframe contain now?

```
NYCPay_df <- NYC_Pay %>%
  filter(Fiscal.Year == 2015, Work.Location.Borough == 'MANHATTAN')
```

Comments: The dataframe now has 405,339 rows and 17 columns.

Task 4 (a): A. Call the str() function on your dataset. As what data type does R see the variable Agency.Start.Date? If this variable is not seen as a date, fix this – turn this into a "Date" data type.

```
str(NYCPay_df$Agency.Start.Date)
```

```
##  chr [1:405339] "04/08/2013" "12/06/1993" "07/19/2010" "06/17/2013" ...
```

```
NYCPay_df$Agency.Start.Date <- as.Date(NYCPay_df$Agency.Start.Date, '%m/%d/%Y')

str(NYCPay_df$Agency.Start.Date)
```

```
##  Date[1:405339], format: "2013-04-08" "1993-12-06" "2010-07-19" "2013-06-17" "1979-11-20" ...
```

Task 4 (b): Create a new variable in the dataframe called Longevity. Longevity should be the number of days between the last date in your fiscal year, and a person's agency start date.

```
NYCPay_df$Longevity <- as.Date('2015-06-30') - NYCPay_df$Agency.Start.Date
```

Task 4 (c): Use the arrange() function from dplyr to sort the employees by longevity, in descending order.

```
NYCPay_df1 <- NYCPay_df %>%
  filter(Agency.Start.Date != '9999-12-31') %>%
  arrange(desc(Longevity))
```

Comment: The longest-serving employee is David Streko in Dept. of Ed Pedagogical as a Teacher.

Task 5: Remove the variable Payroll.Number

```
NYCPay_df2 <- subset(NYCPay_df, select = -(Payroll.Number))
```

Task 6: Create a feature called Annual.Pay. It should be the sum of Regular.Gross.Paid + Total.OT.Paid + Total.Other.Pay.

```
NYCPay_df3 <- NYCPay_df2 %>%
  mutate(Annual.Pay = (Regular.Gross.Paid + Total.OT.Paid + Total.Other.Pay))
```

Task 7: Identify the 8 most common agencies in the dataframe

```
Top8Agency <- as.data.frame(sort(table(NYCPay_df3$Agency.Name), decreasing = TRUE)[1:8])
```

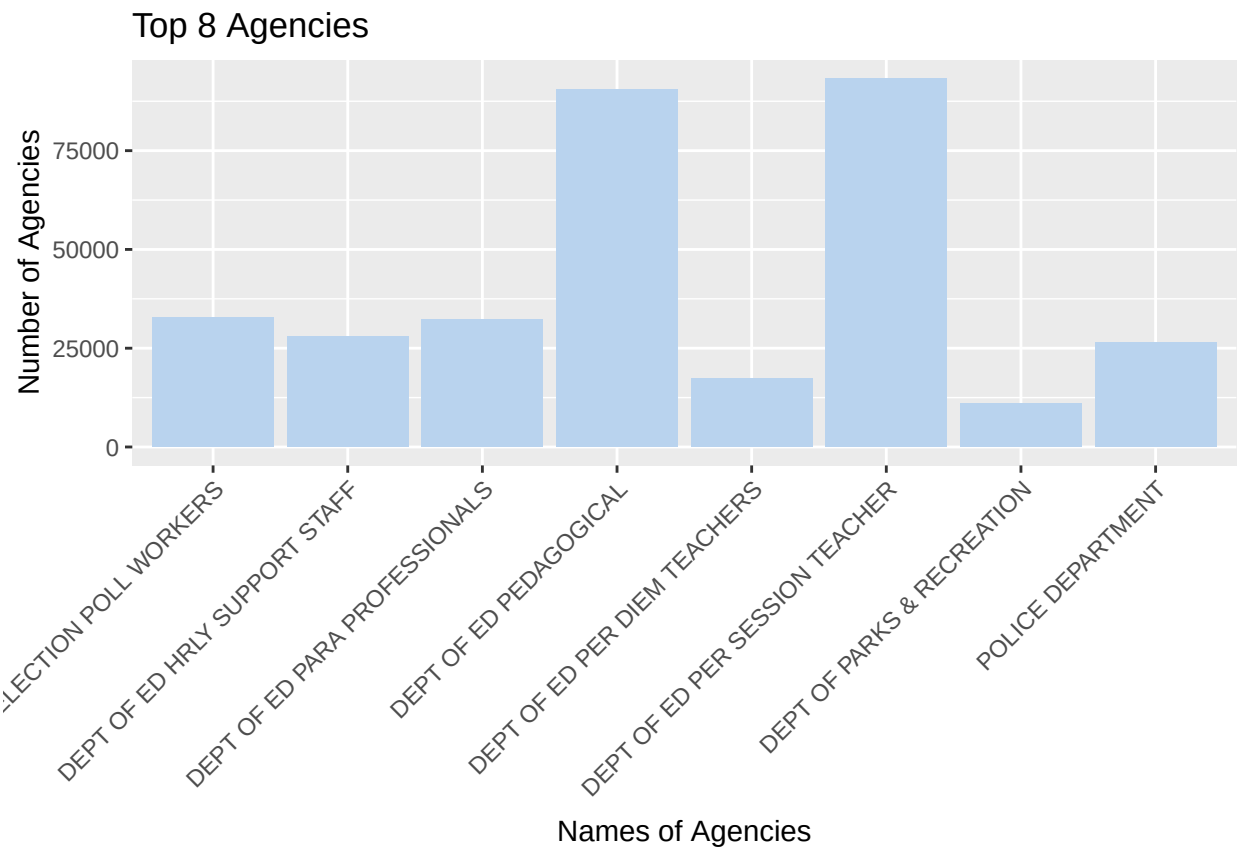
Comments: The 8 most common agencies are Dept. of Ed Per Session Teacher, Dept. of Ed Pedagogical, Board of Election Poll Workers, Dept. of Ed Para Professionals, Dept. of Ed Hrly Support Staff, Police Department, Dept. of Ed Per Diem Teachers and Dept. of Parks and Recreation.

Task 8: Create a new dataframe that only contains data for the eight most common agencies for your year & borough

```
NYCPay_df4 <- filter (NYCPay_df3, Agency.Name %in% Top8Agency$Var1)
```

Task 9: Using ggplot, create a barplot that displays the number of records for the eight most-common agencies.

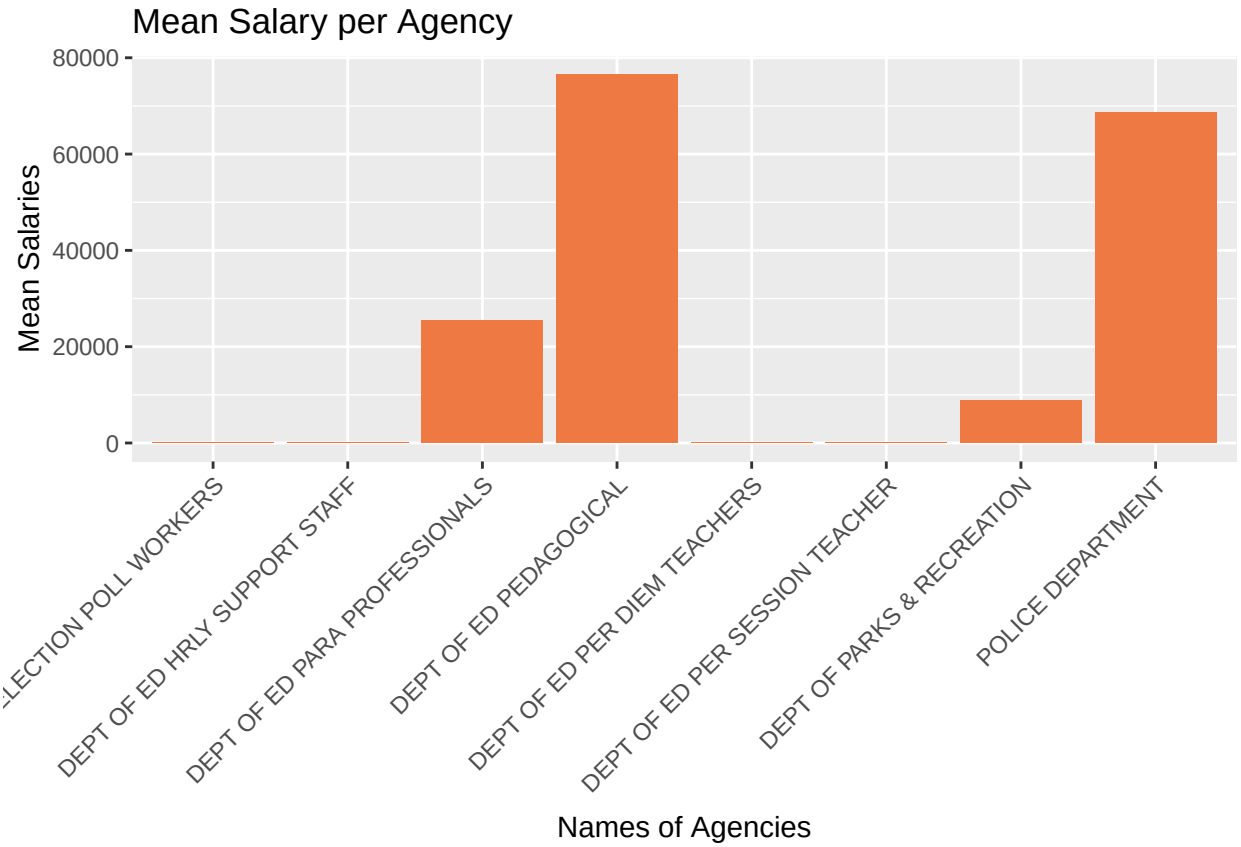
```
ggplot(NYCPay_df4, aes(x = Agency.Name)) +
  geom_bar(fill = 'slategray2') +
  theme(axis.text.x=element_text(angle=45, hjust=1, vjust=1)) +
  ggtitle("Top 8 Agencies") +
  xlab("Names of Agencies") +
  ylab("Number of Agencies")
```




Comments: The bar plot depicts the 8 most common agencies that have employed people in FY 2015 in Manhattan, New York City. It can be seen that the Dept of Ed Pedagogical and Dept of Ed per Session Teacher has hired the maximum employees followed by Dept of Ed Para Professionals and Board of Election Poll Workers.

Task 10: Now let's create another barplot.Put the agency names on one axis, and put the mean salaries per agency on the other axis.

```
ggplot(NYCPay_df4, aes(x = Agency.Name, y = Base.Salary)) +
  stat_summary(fun = 'mean', geom = 'bar', fill = 'sienna2')+
  theme(axis.text.x=element_text(angle=45, hjust=1, vjust=1)) +
  ggtitle("Mean Salary per Agency") +
  xlab("Names of Agencies") +
  ylab("Mean Salaries")
```



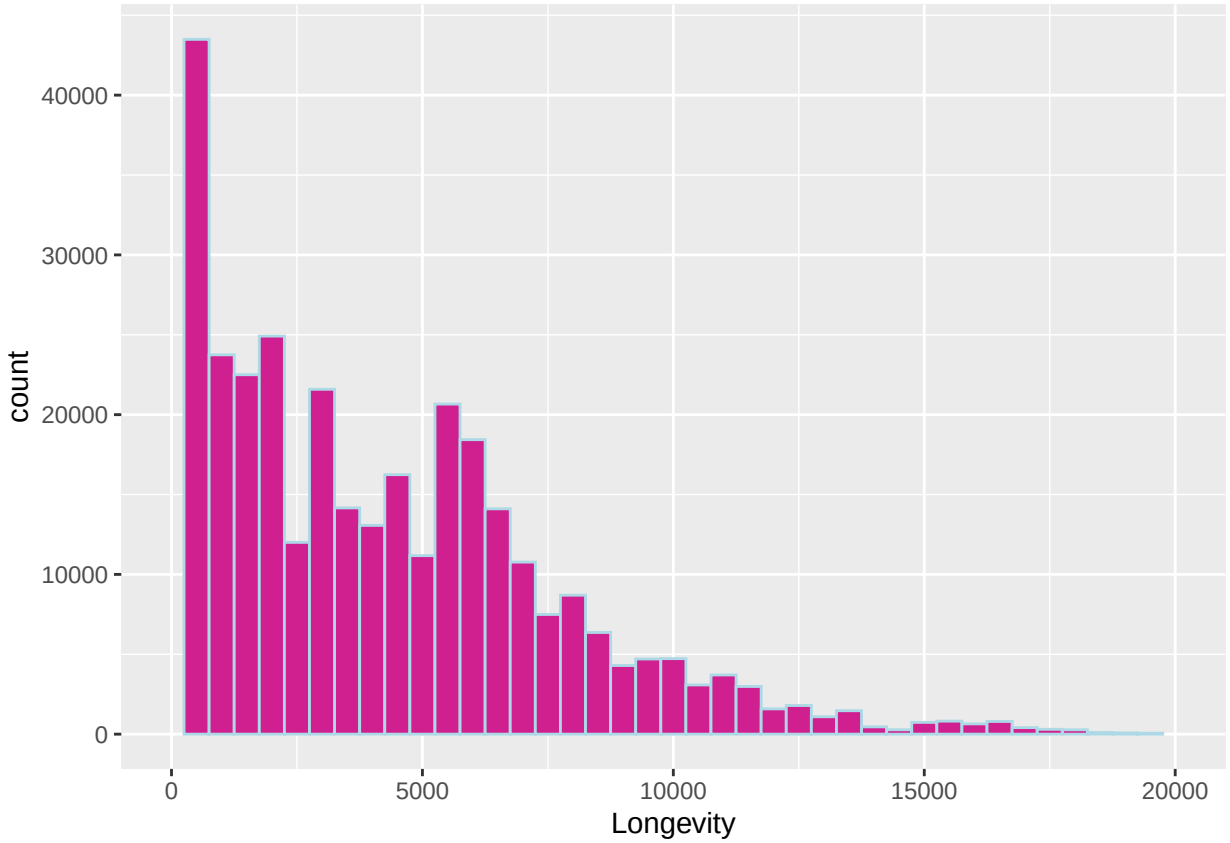

Comments: The barplot shows mean salaries for only 4 agencies which is absurd. According to this plot, Dept of Ed Pedagogical and Police Dept have the highest mean salaries which can be misleading given the fact the barplot lacks salary information of other agencies.

Task 11: Create a histogram that depicts the distribution of the longevity variable for your dataset.

```
Histogram <- ggplot(data = NYCPay_df4, aes(x = Longevity)) +
  geom_histogram(binwidth = 500, fill = 'violetred', color = 'lightblue') +
  xlim(0,20000)
Histogram
```

```
## Warning: Removed 383 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```



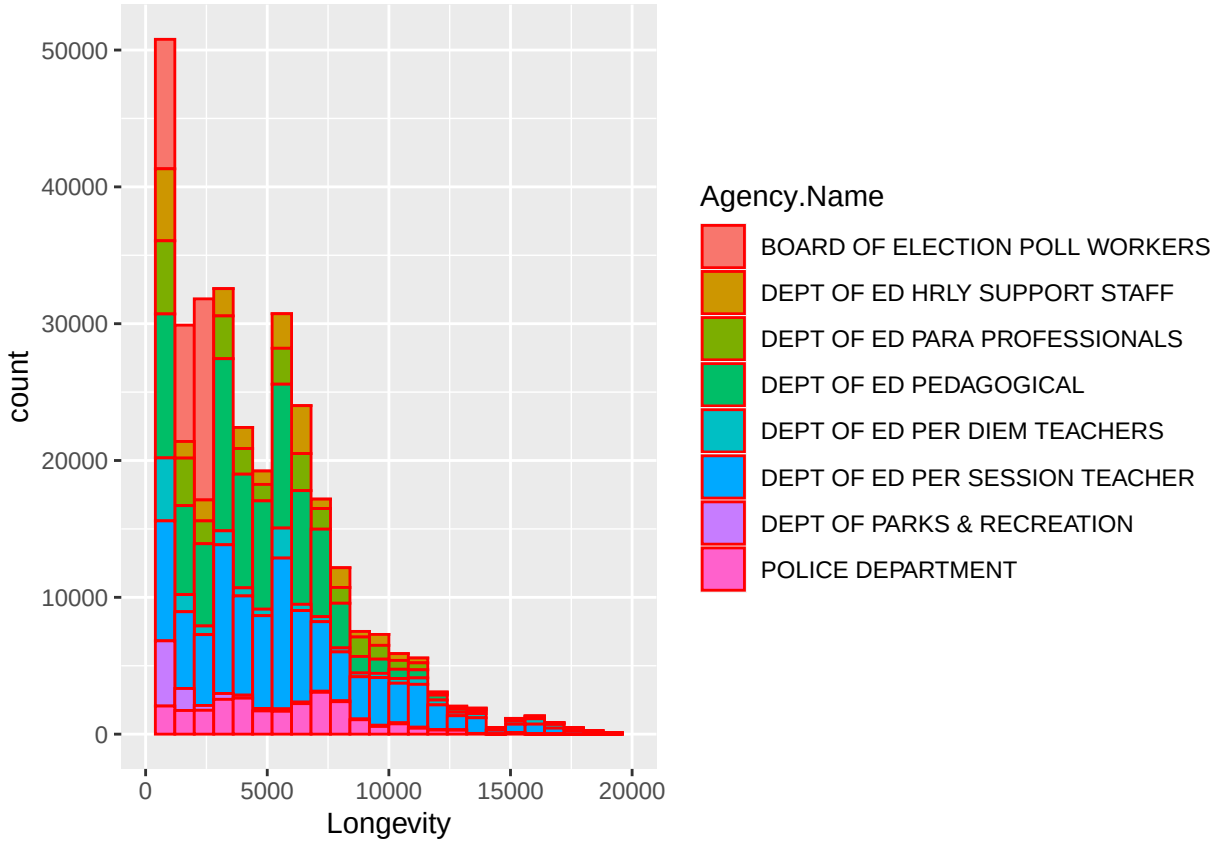

Comment:For this histogram, my data had outliers. My data had a lot of negative values for a particular date '9999-12-31' which seemed unusual and hence I filterd it out. The histogram looks right-skewed indicating a positive skewness and a mean greater than the median.

Task 12: Build another histogram from the previous step, but this time, set the fill parameter in the aesthetics layer to the agency name.

```
ggplot(data = NYCPay_df4, aes(x = Longevity, fill = Agency.Name)) +
  geom_histogram(binwidth = 800, color = 'red') +
  xlim(0,20000)
```

```
## Warning: Removed 383 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 16 rows containing missing values (geom_bar).
```




```
  theme(axis.text.x=element_text(angle=45, hjust=1, vjust=1)) +
  ggtitle("Longevity by Agency Name") +
  xlab("Longevity") +
  ylab("Count")
```

```
## List of 4
##  $ axis.text.x:List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : num 1
##   ..$ vjust        : num 1
##   ..$ angle        : num 45
##   ..$ lineheight   : NULL
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi FALSE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ title      : chr "Longevity by Agency Name"
##  $ x          : chr "Longevity"
##  $ y          : chr "Count"
##  - attr(*, "class")= chr [1:2] "theme" "gg"
##  - attr(*, "complete")= logi FALSE
##  - attr(*, "validate")= logi TRUE
```

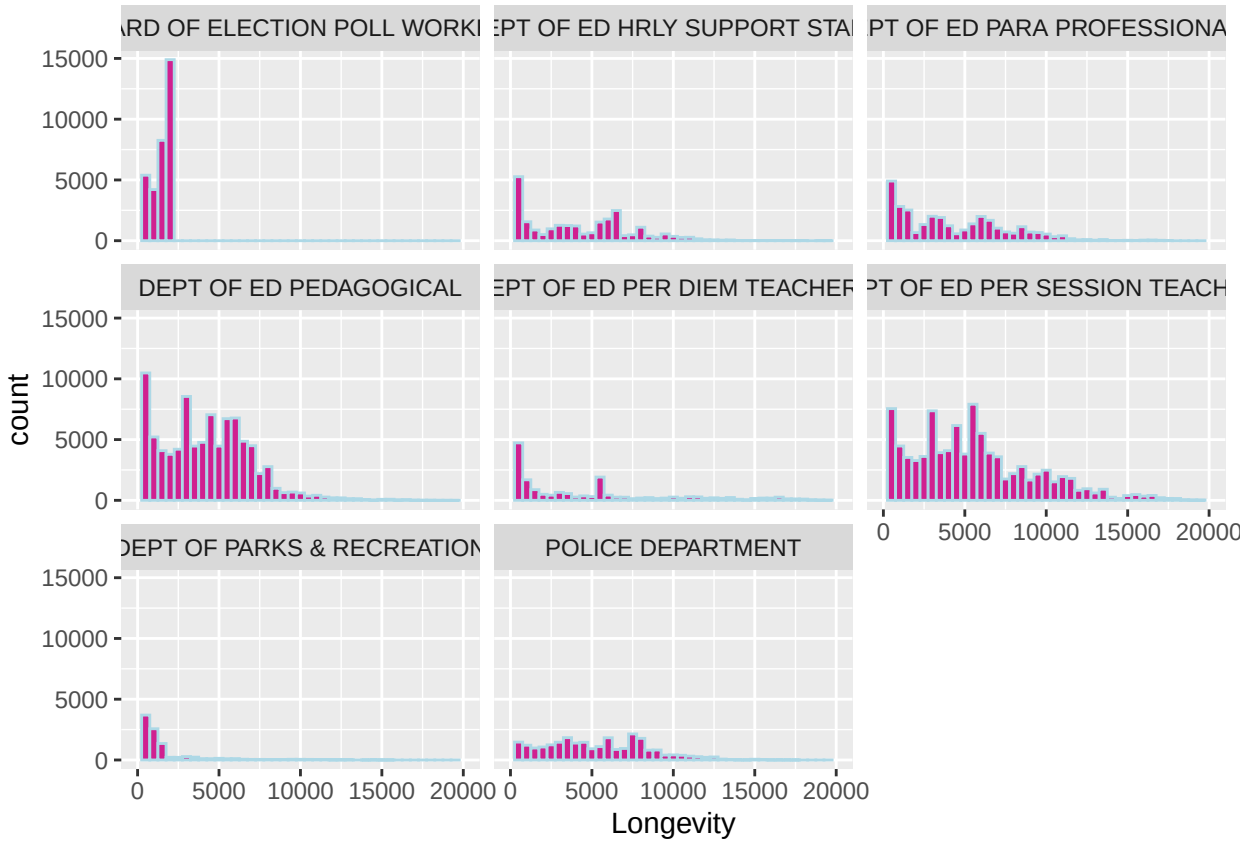
Comment: This histogram gives a clearer view of which agency has longer serving employees by including agency names in the fill() function. It can be seen that longevity for Board of Election Poll Workers ranges from 0 - 5000 days whereas Dept of Ed per Session Teacher and Dept of Ed Pedagogical have higher longevity compared to other agencies.

Task 13: Generate a visualization that lets us separately see the entire longevity histograms for each agency. Use facet_wrap to generate 8 separate histograms within the same plot.Facet on the Agency.Name variable, and use colors of your choice for the fill and borders of the bars.

```
Histogram + facet_wrap(~ Agency.Name)
```

```
## Warning: Removed 383 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 16 rows containing missing values (geom_bar).
```

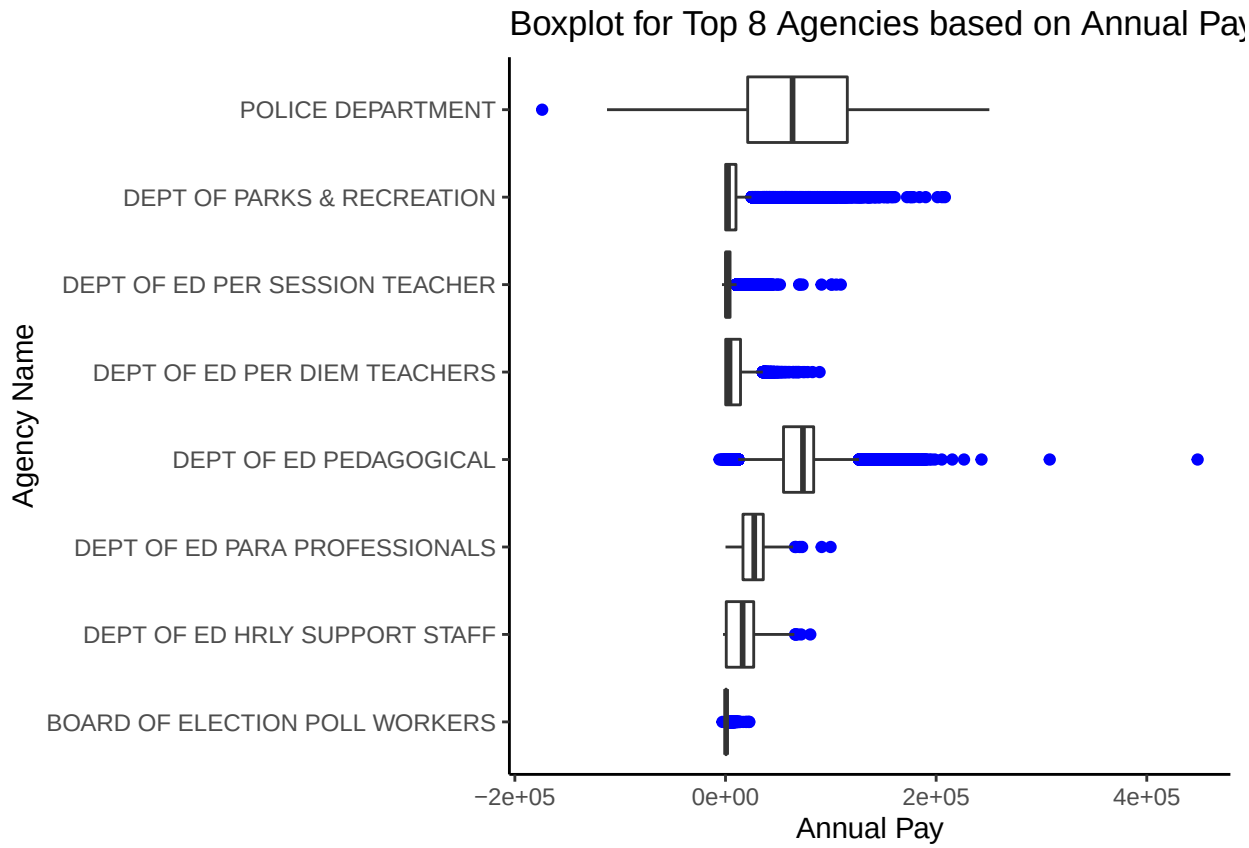



Comment: This histogram gives a holistic view of longevity in each department. This is better than the previous visualization since it is easier to compare the performance over different departments which will help in making better data-driven decisions. Also, much more information can be gained from this histogram about the employees and their longevity in the agencies. For instance, it can be seen that Dept of Ed per Session Teacher and Dept of Ed Pedagogical have higher longevity whereas Dept of Parks and Recreastion has the lowest.

Task 14: Generate a boxplot that depicts agency name on one axis, and total pay on the other.

```
ggplot(data=NYCPay_df4, aes(x= Agency.Name, y=Annual.Pay)) +
  geom_boxplot(outlier.colour = 'blue') +
```

```
  theme_classic() +
  coord_flip() +
  ggtitle("Boxplot for Top 8 Agencies based on Annual Pay") +
  xlab("Agency Name") +
  ylab("Annual Pay")
```



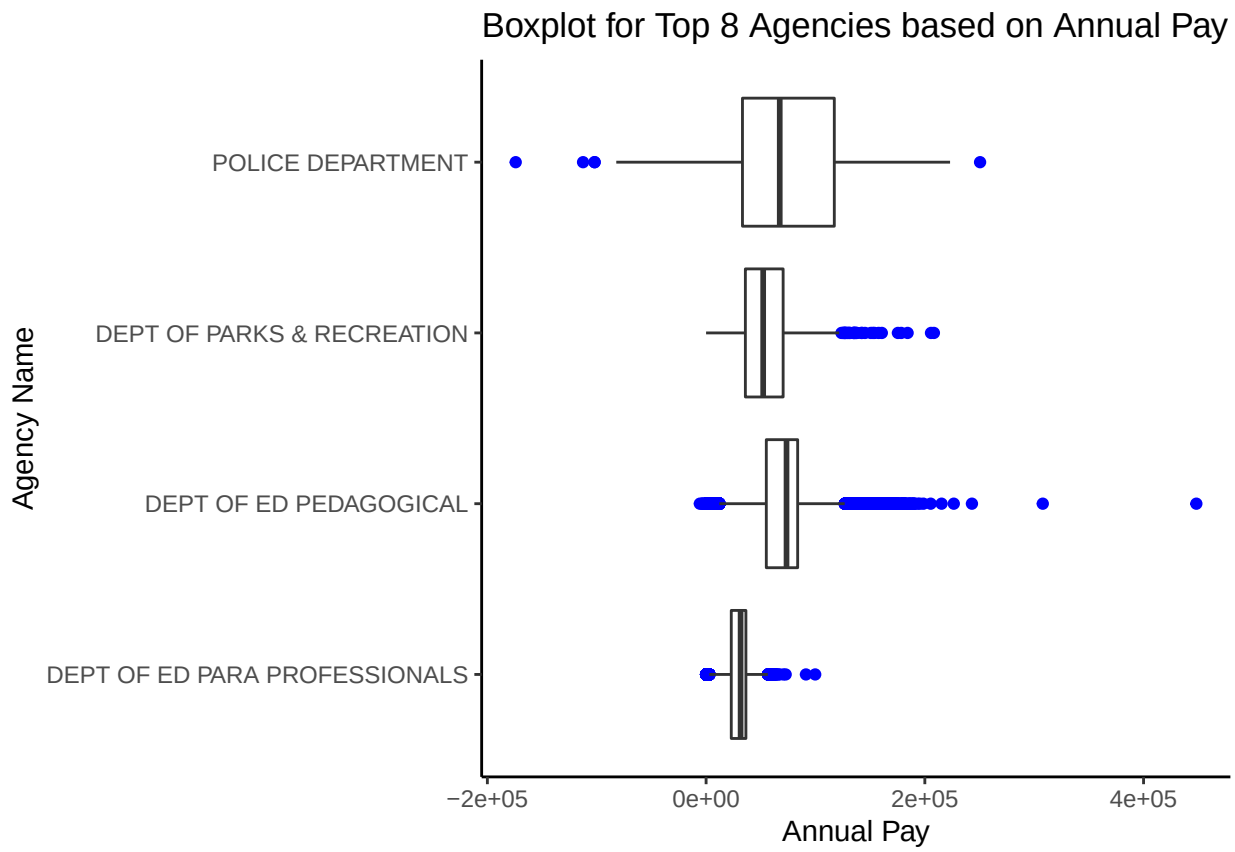

Comment: 1. The above boxplot displays the distribution of the data based on the five number summary (minimum, first quartile (Q1), median, third quartile (Q3) and maximum) along with depicting the outliers in the data. 2. Skewness in a boxplot is when the median cuts the box into unequal pieces. Here, agencies Dept of Ed Pedagogical, Dept of Ed Para Professionals and Dept of Hrly Support Staff are left-skewed indicating that the mean is smaller than the median. Police Dept, Dept of Parks and Recreation and Dept of Ed per Diem Teachers showe a right-skewed data indicating a wider range in the data. Dept of Ed per Session Teacher and Board of Election Poll Workers have symmetric distributions. 3. The boxplot also highlights outliers which are values numerically distant from rest of the data. Here, Dept of Parks and Recreation, Dept of Ed Pedagogical and Dept of Ed per Session Teacher have a significant amount of outliers indicating an experimental error or variability in measurement.It can be due to some senior employees getting paid more than other employees.

Task 15: Filter the dataframe that you are currently using so that only records with a pay basis of "Per Annum" are included. Now, re-create the boxplot that you built in the previous step.

```
NYCPay_df5 <- NYCPay_df4 %>%
  filter(Pay.Basis == 'per Annum')

ggplot(data=NYCPay_df5, aes(x= Agency.Name, y=Annual.Pay)) +
  geom_boxplot(outlier.colour = 'blue') +
  theme_classic() +
```

```
  coord_flip() +
  ggtitle("Boxplot for Top 8 Agencies based on Annual Pay") +
  xlab("Agency Name") +
  ylab("Annual Pay")
```




```
table(NYCPay_df4$Pay.Basis, NYCPay_df4$Agency.Name)
```

```
##
##                  BOARD OF ELECTION POLL WORKERS DEPT OF ED HRLY SUPPORT STAFF
##   per Annum                                   0                             0
##   per Day                                     0                             0
##   per Hour                                32787                         28083
##   Prorated Annual                             0                             0
##
##                  DEPT OF ED PARA PROFESSIONALS DEPT OF ED PEDAGOGICAL
##   per Annum                              25766                  90475
##   per Day                                 6437                      0
##   per Hour                                  56                      0
##   Prorated Annual                            0                      0
##
##                  DEPT OF ED PER DIEM TEACHERS DEPT OF ED PER SESSION TEACHER
##   per Annum                                 0                              0
##   per Day                               17372                          93342
##   per Hour                                  0                              0
##   Prorated Annual                           0                              0
```

```
##
##                   DEPT OF PARKS & RECREATION POLICE DEPARTMENT
##   per Annum                             1737             25240
##   per Day                                144               179
##   per Hour                              9129              1125
##   Prorated Annual                          0                 7
```

Comment: This boxplot has been filtered by "pay per annum" which has removed 4 agencies since their pay was based on per day, per hour or prorated annual. As it can be seen, out of the top 8 agencies, 2 pay their employees only per hour, 2 pay only per day and the remaining 4 pay per annum as well as by day / hour / prorated annual. Here too, Dept of Pedagocial has the highest outliers. Moreover, Dept of Parks and Recreation shows a symmetric distribution.