

CIS 5250

NYC Payroll Data Analysis Using R



Submitted To:

Dr. Shilpa Balan

Submitted By:

Shailja Pandit

CIN: 401971060

Pooja Madhup

CIN: 401977573

Contents

| | |
|--|----|
| Introduction..... | 3 |
| Data Description..... | 4 |
| Data Cleaning..... | 10 |
| Analysis and Visualizations..... | 21 |
| Statistical Summary and Functions..... | 31 |
| Conclusion..... | 37 |
| Reference..... | 37 |

INTRODUCTION

The New York State Payroll Citywide Dataset is a valuable resource for individuals and organizations interested in understanding the salaries and compensation of city employees in New York. This dataset provides detailed information on the salaries of more than 300,000 city employees, including their job titles, departments, and pay rates. By analyzing this dataset, researchers and policymakers can gain insights into the workings of the city government, identify potential areas for cost savings, and ensure that city employees are being fairly compensated for their work. Additionally, this dataset can be used to study trends in the labor market and to evaluate the effectiveness of different policies and practices. Overall, the New York State Payroll Citywide Dataset is an important tool for understanding the inner workings of the city and improving its operations.

This Data is collected because of public interest in how the City's budget is being spent on salary and overtime pay for all municipal employees. This data is provided by the Office of Payroll Administration (OPA). This data will be used by students/recent graduates and other job seekers who are yet to decide whether to apply for public or private sector jobs. This data caters most of the public sector departments. The public sector jobs come with its own set of benefits such as added job security, health benefits and lucrative pension benefits. This dataset will help an individual to decide which department to choose from specifically based on the location, salary, and overtime (OT) hours. According to the article (Eliza Sayon et al. 2021) the five main benefits of working for a private sector firm are Job Security, Employee Benefits, Training, Aiding the community and Better Job-related opportunities. According to the Springer Link article (Jonathan Brock et al. 2001) the public sector in the US employees approximately 14.5

percent of the workforce and the Public Sector Employment is further divided into three categories which are Federal, State and Local Government.

Motivation: This data can be used to analyze how the City's financial resources are allocated and how much of the City's budget is being devoted to overtime. The reader of this data should be aware that increments of salary increase received over the course of any one fiscal year will not be reflected. All that is captured is the employee's final base and gross salary at the end of the fiscal year.

The goal of this project is to provide insights about the payroll activities in the New York to understand users / employees' expectations better and thus help coming employees and current working staff to get clarity regarding financial section of the New York state in the America.

DATASET INFORMATION

Source of Data: Kaggle.com [NY Citywide Payroll Data \(Fiscal Year\) | Kaggle](#)

Data Description:

Data is input into the City's Personnel Management System ("PMS") by the respective user agencies. Each record represents the following statistics for every city employee: Agency, Last Name, First Name, Middle Initial, Agency Start Date, Work Location Borough, Job Title Description, Leave Status as of the close of the FY (June 30th), Base Salary, Pay Basis, Regular Hours Paid, Regular Gross Paid, Overtime Hours worked, Total Overtime Paid, and Total Other Compensation (i.e. lump sum and/or retro payments). This data can be used to analyze how the City's financial resources are allocated and how much of the City's budget is being devoted to

overtime. Increments of salary increases received over the course of any one fiscal year will not be reflected. All that is captured, is the employee's final base and gross salary at the end of the fiscal year.

Below listed are all the field names and the corresponding description for each field name. An additional column of Sample value is also added for the viewer to get an idea what can be expected in the respective fields.

| Field Name | Field Description | Sample Value |
|--------------------------|--|--|
| Fiscal year | Financial year | 2020, 2018 |
| Payroll Number | The payroll agency that the employee works for | 67, 3,11 |
| Agency Name | Agency Names of Municipal employees | ADMIN FOR CHILDREN'S SVCS, DEPARTMENT OF BUILDINGS |
| Last Name | Last name of the employee | MAKHRINSKY, SHAH |
| First Name | First name of the employee | HETAL, LINDA |
| Middle Initial | The middle initial of employee | P, A, C, M |
| Agency Start Date | The date which employee began working for their current agency | 2011-10-03T00:00:00.000 |
| Work Location Borough | Borough of employee's primary work location | MANHATTAN, BRONX |
| Title Description | Civil service title description of the employee | STAFF ANALYST, ADMINISTRATIVE |

| | | |
|---------------------------|--|------------------------------|
| | | DIRECTOR OF SOCIAL SERVICES |
| Leave Status as of Jun 30 | Status of employee as of the close of the relevant fiscal year: Active, Ceased, or On Leave | ACTIVE, SEASONAL, ON LEAVE |
| Base Salary | Base Salary assigned to the employee. Base Salary represents the amount the job pays (not necessarily what was earned) and not including any other pay (differentials, lump sums, uniform allowance, meal allowance, retroactive pay increases, settlement amounts, etc.) or overtime | 86096, 67868, 18 |
| Pay Basis | Lists whether the employee is paid on an hourly, per diem or annual basis | per Day, per Annum, per Hour |
| Regular Hours | Number of regular hours employee worked in the fiscal year . This does not include overtime hours | 1820, 80 |

| | | |
|---------------------|--|----------------|
| Regular Gross Paid | <p>The amount paid to the employee for base salary during the fiscal year.</p> <p>Regular gross paid represents actual base salary during reporting period, which is the portion of the person's annual salary paid before deductions are calculated withheld. This does not include overtime pay or other compensation and does not reflect the after-tax amount or net pay.</p> <p>Total gross pay is calculated by adding columns L, N and O.</p> | 89370.34, 38.4 |
| OT Hours (Overtime) | Overtime Hours worked by employee in the fiscal year. OT stands for Overtime | 5-hour, 7-hour |
| Total OT Paid | Total overtime pay paid to the employee in the fiscal year. OT stands for Overtime | 300, 277 |
| Total Other Pay | Includes any compensation in addition to gross salary and overtime pay, i.e., differentials, lump sums, uniform allowance, meal allowance, retroactive pay increases, settlement amounts, and bonus pay, if applicable. Not every employee will have a value in this field. | 798.8, 817.55 |

| | | |
|--|--|--|
| | For those employees with no other pay, earnings will be stated as \$0 | |
|--|--|--|

Before we clean the data, we read csv dataset in the RStudio using readr library as we know 'readr' library is used for reading, writing tabular data such as CSV file and view the dataset in the next tab in Rstudio using view () function. We also install some packages and run libraries such as 'tidyverse' is a collection of packages, which are designed to work together to make data manipulation and visualization easier in R, 'dplyr' is a package that provides a set of functions for manipulating data frames. It is also part of the tidyverse and is often used with readr to import and clean up data. Another library is 'scales' is a package that provides functions for formatting numeric values and creating visualizations, such as adding labels to plots or setting the limits of axes. It is also part of the tidyverse. Also, 'RColorBrewer' is a package that provides a set of predefined color palettes for use in R plots. It can be used with scales to set the color of plotted data points or lines. and soon to fulfill our objective as to analyze with different use cases and better understand using visualized graphs/charts

Screenshot of Data:

We have 17 total columns in our dataset and 14 usable columns. Here is the code and screenshot of the dataset.

```
> library(tidyverse)
```



```

> library(dplyr)

> library(tidyr)

> library(scales)

> library(RColorBrewer)

> library(readr)

> NYCPayroll_Data <- read_csv("NYCPayroll_Data.csv")

> View(NYCPayroll_Data)

```

Console

Terminal

Background Jobs

R 4.2.2 · ~/4th semester CIS/CIS 5250 Visual Analytics/R Project/

```

> library(tidyverse)
> library(dplyr)
> library(tidyr)
> library(scales)
> library(RColorBrewer)
>
> library(readr)
> NYCPayroll_Data <- read_csv("NYCPayroll_Data.csv")
Rows: 632666 Columns: 17— Column specification
Delimiter: ","
chr (9): AgencyName, LastName, FirstName, MidInit, WorkLocationBorough, Tit...
dbl (7): FiscalYear, PayrollNumber, RegularHours, RegularGrossPaid, OTHours...
dtm (1): AgencyStartDate
i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
> View(NYCPayroll_Data)

```

| FiscalYear | PayrollNumber | AgencyName | LastName | FirstName | MidInit | AgencyStartDate | WorkLocationBorough | TitleDescription | LeaveStatusasofJune30 | BaseSalary | PayBasis | RegularHours | RegularGrossPaid | OTHours | TotalOTPaid | TotalOtherPay |
|------------|---------------|---------------------------|-----------------|-----------|---------|-----------------|---------------------|--|-----------------------|------------|-----------|--------------|------------------|---------|-------------|---------------|
| 2020 | 87 | ADMIN FOR CHILDREN'S SVCS | MAKHINSKY | IRINA | NA | 2011-10-03 | MANHATTAN | STAFF ANALYST | ACTIVE | \$81,509 | per Annum | 1820.00 | 6000.00 | 0 | 0 | 0 |
| 2020 | 87 | ADMIN FOR CHILDREN'S SVCS | SANTOS | WESLEY | NA | 2017-02-21 | MANHATTAN | PROGRAM EVALUATOR | ACTIVE | NA | per Annum | 1820.00 | 79325.99 | 0 | 0 | 0 |
| 2020 | 87 | ADMIN FOR CHILDREN'S SVCS | SLAREZ | JOSEPHINE | NA | 2001-05-21 | MANHATTAN | DIRECTOR OF FIELD OPERATIONS | ACTIVE | \$95,891 | per Annum | 1820.00 | 94379.02 | 0 | 0 | 0 |
| 2020 | 87 | ADMIN FOR CHILDREN'S SVCS | SEALEY | RHONDA | P | 2015-06-01 | MANHATTAN | ADMINISTRATIVE DIRECTOR OF SOCIAL SERVICES | ACTIVE | \$97,755 | per Annum | 1820.00 | 96354.13 | 0 | 0 | 0 |
| 2020 | 87 | ADMIN FOR CHILDREN'S SVCS | SEALEY | RHONDA | P | 2015-06-01 | MANHATTAN | ADMINISTRATIVE DIRECTOR OF SOCIAL SERVICES | ACTIVE | \$97,755 | per Annum | 1820.00 | 96354.13 | 0 | 0 | 0 |
| 2020 | 87 | ADMIN FOR CHILDREN'S SVCS | SMITHSON | FREDIE | NA | 2000-11-20 | MANHATTAN | PROCUREMENT ANALYST | ACTIVE | \$71,706 | per Annum | 1820.00 | 71318.40 | 0 | 0 | 0 |
| 2020 | 87 | ADMIN FOR CHILDREN'S SVCS | RODRIGUEZ | EDWIN | J | 2018-12-10 | MANHATTAN | COMMUNITY COORDINATOR | ACTIVE | \$71,380 | per Annum | 1820.00 | 70264.54 | 0 | 0 | 0 |
| 2020 | 87 | ADMIN FOR CHILDREN'S SVCS | LEBAKIS | LUZ | M | 2013-10-28 | MANHATTAN | COMMUNITY COORDINATOR | CEASED | 65562 | per Annum | 70.00 | 2511.77 | 0 | 0 | 0 |
| 2020 | 87 | ADMIN FOR CHILDREN'S SVCS | ISOLA | MICHAEL | J | 2011-04-25 | MANHATTAN | COMPUTER SYSTEMS MANAGER | ACTIVE | 140289 | per Annum | 1820.00 | 128071.19 | 0 | 0 | 0 |
| 2020 | 87 | ADMIN FOR CHILDREN'S SVCS | MEDLEY | LESLIE | F | 1996-06-23 | MANHATTAN | DIRECTOR OF FIELD OPERATIONS | ACTIVE | 113956 | per Annum | 1820.00 | 109364.80 | 0 | 0 | 0 |
| 2020 | 87 | ADMIN FOR CHILDREN'S SVCS | BREWINGTON | PAULA | M | 2014-11-24 | MANHATTAN | COMMUNITY COORDINATOR | ACTIVE | 73010 | per Annum | 1820.00 | 72079.50 | 0 | 0 | 0 |
| 2020 | 87 | ADMIN FOR CHILDREN'S SVCS | MOYE | COHEN | L | 2015-10-09 | MANHATTAN | COMMUNITY COORDINATOR | ACTIVE | 73560 | per Annum | 1820.00 | 70264.54 | 0 | 0 | 0 |
| 2020 | 87 | ADMIN FOR CHILDREN'S SVCS | HUBER | RONALD | J | 2014-03-03 | MANHATTAN | ADMINISTRATIVE STAFF ANALYST | ACTIVE | 76835 | per Annum | 1820.00 | 76419.64 | 0 | 0 | 0 |
| 2020 | 87 | ADMIN FOR CHILDREN'S SVCS | YUSAF | SALISHA | S | 2007-01-29 | MANHATTAN | ADMIN COMMUNITY RELATIONS SPECIALIST | ACTIVE | 79015 | per Annum | 1820.00 | 78587.92 | 0 | 0 | 0 |
| 2020 | 87 | ADMIN FOR CHILDREN'S SVCS | RODRIGUEZ CORRE | IVELISSE | NA | 2013-03-11 | MANHATTAN | ADMINISTRATIVE STAFF ANALYST | ACTIVE | 82500 | per Annum | 1820.00 | 82054.18 | 0 | 0 | 0 |
| 2020 | 87 | ADMIN FOR CHILDREN'S SVCS | BENT | DENISE | M | 2014-04-28 | MANHATTAN | CHILD AND FAMILY SPECIALIST | ACTIVE | 83981 | per Annum | 1820.00 | 82991.66 | 0 | 0 | 0 |
| 2020 | 87 | ADMIN FOR CHILDREN'S SVCS | HUTCHINSON | IAN | W | 2016-07-18 | MANHATTAN | DIRECTOR OF FIELD OPERATIONS | ACTIVE | 87560 | per Annum | 1820.00 | 86219.78 | 0 | 0 | 0 |
| 2020 | 87 | ADMIN FOR CHILDREN'S SVCS | BLUE | WYNNE | J | 2019-02-25 | BROOKLYN | YOUTH DEVELOPMENT SPECIALIST | CEASED | 40739 | per Annum | 253.80 | 6338.58 | 0 | 0 | 0 |
| 2020 | 87 | ADMIN FOR CHILDREN'S SVCS | LU | YING | M | 2018-12-24 | MANHATTAN | ASSOCIATE STAFF ANALYST | CEASED | 60751 | per Annum | 70.00 | 2319.24 | 0 | 0 | 0 |
| 2020 | 87 | ADMIN FOR CHILDREN'S SVCS | AHLGREN | SOLA | A | 2006-05-01 | MANHATTAN | ADMINISTRATIVE STAFF ANALYST | ACTIVE | 100000 | per Annum | 1750.00 | 99459.62 | 0 | 0 | 0 |

DATA CLEANING

Data cleaning is an important step before data visualization because it helps ensure that the data is in a usable format and contains accurate and reliable values. Data cleaning involves a variety of tasks, such as removing missing values, converting data types, and formatting values. By performing data cleaning before data visualization, we can ensure that our dataset is ready for analysis and that the visualizations you create are based on accurate and reliable data. This can help us to avoid common problems, such as errors and incorrect results, that can occur when working with dirty data. It can also help improve the quality and effectiveness of your visualizations. We used different types of data cleaning techniques to clean our dataset which are as follows:

1. Removing NA

Before Cleaning:

Removing NA and blank values from a dataset as part of the data cleaning process is important because it helps to ensure that our dataset is easier to understand the trends and patterns while doing visualization. It can improve the clarity and interpretability of the results and can reduce the size of the dataset, which can be useful if the dataset is very large and is taking up a lot of memory to improve the performance of data. Additionally, removing NA and blank values can help to ensure that our dataset is complete and accurate. Overall, removing these values is an essential step in the data cleaning process that can help to improve the quality and reliability of your analysis. There were some 'NA' values in the Midinit and BaseSalary columns as shown in the screenshot. Before removing NA value, first we check the dimension of the dataset using 'dim' function, then we find out the sum of total count of NA values present in our dataset.

Code:

```
# Check the total number of missing values in the data set
```

```
sum(is.na (NYCPayroll_Data))
```

```
> dim(NYCPayroll_Data)
[1] 632666    17
> sum(is.na(NYCPayroll_Data))
[1] 476138
```

Screenshots:

| MidInit | AgencyStartDate | WorkLocationBorough | TitleDescription | LeaveStatusasofJune30 | BaseSalary |
|---------|-----------------|---------------------|--|-----------------------|------------|
| NA | 2011-10-03 | MANHATTAN | STAFF ANALYST | ACTIVE | \$81,509 |
| NA | 2017-02-21 | MANHATTAN | PROGRAM EVALUATOR | ACTIVE | NA |
| NA | 2001-05-21 | MANHATTAN | DIRECTOR OF FIELD OPERATIONS | ACTIVE | \$95,831 |
| P | 2015-06-01 | MANHATTAN | ADMINISTRATIVE DIRECTOR OF SOCIAL SERVICES | ACTIVE | \$97,755 |
| P | 2015-06-01 | MANHATTAN | ADMINISTRATIVE DIRECTOR OF SOCIAL SERVICES | ACTIVE | \$97,755 |
| NA | 2000-11-20 | MANHATTAN | PROCUREMENT ANALYST | ACTIVE | \$71,706 |
| J | 2018-12-10 | MANHATTAN | COMMUNITY COORDINATOR | ACTIVE | \$71,360 |
| M | 2013-10-28 | MANHATTAN | COMMUNITY COORDINATOR | CEASED | 65562 |
| J | 2011-04-25 | MANHATTAN | COMPUTER SYSTEMS MANAGER | ACTIVE | 140899 |
| F | 1996-06-23 | MANHATTAN | DIRECTOR OF FIELD OPERATIONS | ACTIVE | 113956 |

```
`Payroll Number` `Agency Na...` `Last ...` `First...` `Mid I...` `Agency Start Date` `work ...` `Title...` `Leave...` `Base ...` `Pay B...` `Regul...` `Regul...` `OT`
<dbl> <chr> <chr> <chr> <chr> <dtm> <chr> <chr> <chr> <dbl> <chr> <dbl> <dbl>
67 ADMIN FOR ... MAKHRI... IRINA NA 2011-10-03 00:00:00 MANHAT... STAFF ... ACTIVE 81509 per An... 1820 NA
67 ADMIN FOR ... ACEVEDO MIGUEL... P 2006-02-27 00:00:00 BRONX CHILD ... ACTIVE 86096 per An... 1820 89600.
67 ADMIN FOR ... ARGUETA REGINA A 2015-02-09 00:00:00 BROOKL... CHILD ... ACTIVE 60327 per An... 1820 58938.
67 ADMIN FOR ... ARGUETA REGINA A 2015-02-09 00:00:00 BROOKL... CHILD ... ACTIVE 60327 per An... 1820 58938.
67 ADMIN FOR ... SANTOS WESLEY NA 2017-02-21 00:00:00 MANHAT... PROGRA... ACTIVE NA per An... 1820 79326.
67 ADMIN FOR ... BURKE KEVIN C 2000-03-13 00:00:00 MANHAT... ADMINI... ACTIVE 90764 per An... 1820 89370.
lated variable names: `Agency Name` `Last Name` `First Name` `Mid Title` `Work Location Borough` `Title Description`
```

After Cleaning:

We decided to remove the NA values, so that we get more accuracy in our visualization using “omit” function. Afterwards, we count the total sum of NA values to get confirm that there is no NA value left. The ‘is.na’ function is used to identify the NA values in the dataset, and the ‘sum(is.na)’ function is used to count the number of NA values in the dataset. This can be useful for assessing the quality of the data and determining how many NA values need to be removed.

The omit function is then used to remove these NA values from the dataset. This allows us to easily and efficiently clean our dataset and prepare it for analysis. We viewed the first 6 rows of the dataset after removing NA values. In addition, we used again `sum(is.na())` function to confirm that there was not any countable number of missing values in our dataset. To conclude, we get the productive dataset for analyzing purpose.

Code:

```
#Removing NA and blank data

# Check the total number of missing values in the data set
sum(is.na(NYCPayroll_Data))

# Omit the observations that have missing values
payroll <- na.omit(NYCPayroll_Data)

# Confirm that there are no missing values now
sum(is.na(payroll))

View(payroll)

head(payroll)
```

```
> sum(is.na(NYCPayroll_Data))
[1] 476138
> payroll <- na.omit(NYCPayroll_Data)
> sum(is.na(payroll))
[1] 0
> view(payroll)
> head(payroll)
# A tibble: 6 x 17
```

Screenshots:

| MidInit | AgencyStartDate | WorkLocationBorough | TitleDescription | LeaveStatusasofJune30 | BaseSalary |
|---------|-----------------|---------------------|--|-----------------------|------------|
| P | 2015-06-01 | MANHATTAN | ADMINISTRATIVE DIRECTOR OF SOCIAL SERVICES | ACTIVE | \$97,755 |
| P | 2015-06-01 | MANHATTAN | ADMINISTRATIVE DIRECTOR OF SOCIAL SERVICES | ACTIVE | \$97,755 |
| J | 2018-12-10 | MANHATTAN | COMMUNITY COORDINATOR | ACTIVE | \$71,360 |
| M | 2013-10-28 | MANHATTAN | COMMUNITY COORDINATOR | CEASED | 65562 |
| J | 2011-04-25 | MANHATTAN | COMPUTER SYSTEMS MANAGER | ACTIVE | 140899 |
| F | 1996-06-23 | MANHATTAN | DIRECTOR OF FIELD OPERATIONS | ACTIVE | 113956 |
| M | 2014-11-24 | MANHATTAN | COMMUNITY COORDINATOR | ACTIVE | 75010 |
| L | 2018-10-09 | MANHATTAN | COMMUNITY COORDINATOR | ACTIVE | 71360 |
| J | 2014-03-03 | MANHATTAN | ADMINISTRATIVE STAFF ANALYST | ACTIVE | 76835 |

```
> head payroll
# A tibble: 6 x 17
FiscalYear PayrollNumber AgencyName LastName FirstName MidInit AgencyStartDate WorkLocationBorough TitleDescription LeaveStatusasofJune30 BaseSalary PayBasis RegularHours RegularGrossPaid OTHours TotalOTPaid TotalOtherPay
1 2020 67 ADMIN FOR CHILDREN'S SVCS SEALEY RHONDA P 2015-06-01 00:00:00 MANHATTAN ADMINISTRATIVE DIRECTOR OF SOCIAL SERVICES ACTIVE $97,755 per Annum 1820 36254. 0 0 0
2 2020 67 ADMIN FOR CHILDREN'S SVCS SEALEY RHONDA P 2015-06-01 00:00:00 MANHATTAN ADMINISTRATIVE DIRECTOR OF SOCIAL SERVICES ACTIVE $97,755 per Annum 1820 36254. 0 0 0
3 2020 67 ADMIN FOR CHILDREN'S SVCS RODRIGUEZ EDWIN J 2018-12-10 00:00:00 MANHATTAN COMMUNITY COORDINATOR ACTIVE $71,360 per Annum 1820 10265. 0 0 0
4 2020 67 ADMIN FOR CHILDREN'S SVCS LEGAKIS LUZ M 2013-10-28 00:00:00 MANHATTAN COMMUNITY COORDINATOR CEASED 65562 per Annum 70 2512. 0 0 0
5 2020 67 ADMIN FOR CHILDREN'S SVCS ISOLA MICHAEL J 2011-04-25 00:00:00 MANHATTAN COMPUTER SYSTEMS MANAGER ACTIVE 140899 per Annum 1820 118671. 0 0 0
6 2020 67 ADMIN FOR CHILDREN'S SVCS MEDLEY LESLIE F 1996-06-23 00:00:00 MANHATTAN DIRECTOR OF FIELD OPERATIONS ACTIVE 113956 per Annum 1820 102655. 0 0 0
```

2. Removing duplicate rows

Before Cleaning:

Removing duplicate data is an essential step in the data cleaning process because it helps to ensure that our data is accurate, efficient, and easy to interpret. Duplicate data creates conflicts and confusion while doing data visualization. As we can see from the screenshot below that the data of row 1st and 2nd are duplicate rows with exactly same values throughout the columns. Therefore, we would be removing duplicate values using the code below as before and after cleaning. We used `dim ()` function to check the dimension count of dataset before duplicate data remove. We check the first six rows of dataset using `head ()` function.

Code:

```
#Checking duplicate rows
```

```
head payroll
```

```
dim payroll)
```

```
> head payroll)
> dim payroll)
[1] 240562      17
> |
```

Screenshots:

| FiscalYear | PayrollNumber | AgencyName | LastName | FirstName | MidInit | AgencyStartDate | WorkLocationBorough | TitleDescription | LeaveStatus |
|------------|---------------|---------------------------|-----------|-----------|---------|-----------------|---------------------|--|-------------|
| 2020 | 67 | ADMIN FOR CHILDREN'S SVCS | SEALEY | RHONDA | P | 2015-06-01 | MANHATTAN | ADMINISTRATIVE DIRECTOR OF SOCIAL SERVICES | ACTIVE |
| 2020 | 67 | ADMIN FOR CHILDREN'S SVCS | SEALEY | RHONDA | P | 2015-06-01 | MANHATTAN | ADMINISTRATIVE DIRECTOR OF SOCIAL SERVICES | ACTIVE |
| 2020 | 67 | ADMIN FOR CHILDREN'S SVCS | RODRIGUEZ | EDWIN | J | 2018-12-10 | MANHATTAN | COMMUNITY COORDINATOR | ACTIVE |
| 2020 | 67 | ADMIN FOR CHILDREN'S SVCS | LEGAKIS | LUZ | M | 2013-10-28 | MANHATTAN | COMMUNITY COORDINATOR | CEASED |
| 2020 | 67 | ADMIN FOR CHILDREN'S SVCS | ISOLA | MICHAEL | J | 2011-04-25 | MANHATTAN | COMPUTER SYSTEMS MANAGER | ACTIVE |

After Cleaning:

While cleaning the data set, we used 'unique' function to remove duplicate values from our payroll data frame. This can make it easier to work with the data and can also improve the performance of any statistical that we applied in the further section. It also makes the results of our analysis easier to under. By using the unique function, we removed these duplicate values and ensure that our data is clean and accurate. Additionally, the unique function can also be used to identify and extract unique values from a dataset, which can be useful for exploratory data analysis. Furthermore, we used dim () function to check the dimension of dataset after removed duplicate data. Along with that, we used head () and view () function to view the payroll dataset.

Code:

```
#Removing duplicate Rows
```

```
payroll <- unique(payroll)
```

```
dim(payroll)
```

```
head(payload)
```

```
view(payload)
```

```
> payroll <- unique(payload)
> dim(payload)
[1] 205735    17
> View(payload)
> head(payload)
```

Screenshots:

| FiscalYear | PayrollNumber | AgencyName | LastName | FirstName | Midinit | AgencyStartDate | WorkLocationBorough | TitleDescription | LeaveStatus |
|------------|---------------|---------------------------|------------|-----------|---------|-----------------|---------------------|--|-------------|
| 2020 | 67 | ADMIN FOR CHILDREN'S SVCS | SEALEY | RHONDA | P | 2015-06-01 | MANHATTAN | ADMINISTRATIVE DIRECTOR OF SOCIAL SERVICES | ACTIVE |
| 2020 | 67 | ADMIN FOR CHILDREN'S SVCS | RODRIGUEZ | EDWIN | J | 2018-12-10 | MANHATTAN | COMMUNITY COORDINATOR | ACTIVE |
| 2020 | 67 | ADMIN FOR CHILDREN'S SVCS | LEGAKIS | LUZ | M | 2013-10-28 | MANHATTAN | COMMUNITY COORDINATOR | CEASED |
| 2020 | 67 | ADMIN FOR CHILDREN'S SVCS | ISOLA | MICHAEL | J | 2011-04-25 | MANHATTAN | COMPUTER SYSTEMS MANAGER | ACTIVE |
| 2020 | 67 | ADMIN FOR CHILDREN'S SVCS | MEDLEY | LESLIE | F | 1996-06-23 | MANHATTAN | DIRECTOR OF FIELD OPERATIONS | ACTIVE |
| 2020 | 67 | ADMIN FOR CHILDREN'S SVCS | BREWINGTON | FINA | M | 2014-11-24 | MANHATTAN | COMMUNITY COORDINATOR | ACTIVE |

| A tibble: 6 × 17 | | | | | | | | | |
|------------------|---------------|---------------------------|------------|-----------|---------|-----------------|---------------------|--|--|
| FiscalYear | PayrollNumber | AgencyName | LastName | FirstName | Midinit | AgencyStartDate | WorkLocationBorough | TitleDescription | |
| 2020 | 67 | ADMIN FOR CHILDREN'S SVCS | SEALEY | RHONDA | P | 2015-06-01 | MANHATTAN | ADMINISTRATIVE DIRECTOR OF SOCIAL SERVICES | |
| 2020 | 67 | ADMIN FOR CHILDREN'S SVCS | RODRIGUEZ | EDWIN | J | 2018-12-10 | MANHATTAN | COMMUNITY COORDINATOR | |
| 2020 | 67 | ADMIN FOR CHILDREN'S SVCS | LEGAKIS | LUZ | M | 2013-10-28 | MANHATTAN | COMMUNITY COORDINATOR | |
| 2020 | 67 | ADMIN FOR CHILDREN'S SVCS | ISOLA | MICHAEL | J | 2011-04-25 | MANHATTAN | COMPUTER SYSTEMS MANAGER | |
| 2020 | 67 | ADMIN FOR CHILDREN'S SVCS | MEDLEY | LESLIE | F | 1996-06-23 | MANHATTAN | DIRECTOR OF FIELD OPERATIONS | |
| 2020 | 67 | ADMIN FOR CHILDREN'S SVCS | BREWINGTON | FINA | M | 2014-11-24 | MANHATTAN | COMMUNITY COORDINATOR | |

6 rows | 1-9 of 17 columns

3. Getting consistent value in a column (removing symbols)

Before Cleaning:

It is often necessary to clean data before performing analysis. One common data cleaning task is removing special characters, such as the dollar sign \$, from numeric values. The reason for this is that the dollar sign \$ is a special character that can cause problems when performing numerical operations on data. Another reason to remove the \$ symbol is to ensure that our data is consistent. If some values in the 'BaseSalary' column have the \$ symbol and others do not, this can cause confusion and make it difficult to perform analysis on the data. By removing the \$

symbol, we can ensure that all values in the column are formatted consistently, which can make our data easier to work with. Initially, we used `view()` and `head()` function to verify the presence of \$ symbol.

Code:

```
#Checking unnecessary symbols or alphabets
```

```
View payroll
```

```
head payroll
```

```
> View payroll
> head payroll
```

Screenshots:

| TitleDescription | LeaveStatusasofJune30 | BaseSalary | PayBasis | RegularHours |
|--|-----------------------|------------|-----------|--------------|
| ADMINISTRATIVE DIRECTOR OF SOCIAL SERVICES | ACTIVE | \$97,755 | per Annum | 1820.00 |
| COMMUNITY COORDINATOR | ACTIVE | \$71,360 | per Annum | 1820.00 |
| COMMUNITY COORDINATOR | CEASED | 65562 | per Annum | 70.00 |
| COMPUTER SYSTEMS MANAGER | ACTIVE | 140899 | per Annum | 1820.00 |
| DIRECTOR OF FIELD OPERATIONS | ACTIVE | 113956 | per Annum | 1820.00 |
| COMMUNITY COORDINATOR | ACTIVE | 75010 | per Annum | 1820.00 |

| TitleDescription <chr> | LeaveStatusasofJune30 <chr> | BaseSalary <chr> | PayBasis <chr> | RegularHours <dbl> |
|--|--------------------------------|---------------------|-------------------|-----------------------|
| ADMINISTRATIVE DIRECTOR OF SOCIAL SERVICES | ACTIVE | \$97,755 | per Annum | 1820 |
| COMMUNITY COORDINATOR | ACTIVE | \$71,360 | per Annum | 1820 |
| COMMUNITY COORDINATOR | CEASED | 65562 | per Annum | 70 |
| COMPUTER SYSTEMS MANAGER | ACTIVE | 140899 | per Annum | 1820 |
| DIRECTOR OF FIELD OPERATIONS | ACTIVE | 113956 | per Annum | 1820 |
| COMMUNITY COORDINATOR | ACTIVE | 75010 | per Annum | 1820 |

After Cleaning:

After verified the \$ special character present in some of the values in the BaseSalary column. We decided to remove the symbol using the “gsub” function. The gsub function in R is commonly used for data cleaning tasks, such as removing special characters from numeric values. The gsub function allows us to search for a specified pattern in a string and replace it with a different value. In the case of removing the \$ symbol from numeric values, the gsub function can be used to search for the \$ symbol in a column of data and replace it with an empty string, resulting in a BaseSalary column of data without the \$ symbol. This can be useful for ensuring that the data is in a consistent format and can be used for numerical operations.

After cleaning, payroll is the name of the data frame, BaseSalary is the column name to search and replace in, and ‘gsub("\\\$", "", payroll\$`BaseSalary`)’ searches for the \$ symbol in the BaseSalary column of the payroll data frame and replaces it with an empty string. This results in a column of data without the \$ symbol, which can be used for numerical operations. Afterwards, we viewed the data set and checked the top six rows of the dataset that the \$ special character is removed

Code:

```
#Removing '$' symbol from Base Salary  
payroll$`BaseSalary` = gsub("\\$", "", payroll$`BaseSalary`)  
View(payroll)  
head(payroll)
```

```
> head payroll
> payroll$`BaseSalary` = gsub("\\$", "", payroll$`BaseSalary`)
> view payroll
> head payroll
```

Screenshots:

| TitleDescription | LeaveStatusasofJune30 | BaseSalary | PayBasis | RegularHours | RegularGrossPaid |
|--|-----------------------|------------|-----------|--------------|------------------|
| ADMINISTRATIVE DIRECTOR OF SOCIAL SERVICES | ACTIVE | 97,755 | per Annum | 1820.00 | 96254.13 |
| COMMUNITY COORDINATOR | ACTIVE | 71,360 | per Annum | 1820.00 | 70264.54 |
| COMMUNITY COORDINATOR | CEASED | 65562 | per Annum | 70.00 | 2511.77 |
| COMPUTER SYSTEMS MANAGER | ACTIVE | 140899 | per Annum | 1820.00 | 128671.19 |
| DIRECTOR OF FIELD OPERATIONS | ACTIVE | 113956 | per Annum | 1820.00 | 109584.80 |
| COMMUNITY COORDINATOR | ACTIVE | 75010 | per Annum | 1820.00 | 72375.50 |
| COMMUNITY COORDINATOR | ACTIVE | 71360 | per Annum | 1820.00 | 70264.54 |
| ADMINISTRATIVE STAFF ANALYST | ACTIVE | 76835 | per Annum | 1820.00 | 76419.64 |
| ADMIN COMMUNITY RELATIONS SPECIALIST | ACTIVE | 79015 | per Annum | 1820.00 | 78587.92 |

| A tibble: 6 × 17 | | | | | |
|--|--------------------------------|---------------------|-------------------|-----------------------|---------------------------|
| TitleDescription <chr> | LeaveStatusasofJune30 <chr> | BaseSalary <chr> | PayBasis <chr> | RegularHours <dbl> | RegularGrossPaid <dbl> |
| ADMINISTRATIVE DIRECTOR OF SOCIAL SERVICES | ACTIVE | 97,755 | per Annum | 1820 | 96254.13 |
| COMMUNITY COORDINATOR | ACTIVE | 71,360 | per Annum | 1820 | 70264.54 |
| COMMUNITY COORDINATOR | CEASED | 65562 | per Annum | 70 | 2511.77 |
| COMPUTER SYSTEMS MANAGER | ACTIVE | 140899 | per Annum | 1820 | 128671.19 |
| DIRECTOR OF FIELD OPERATIONS | ACTIVE | 113956 | per Annum | 1820 | 109584.80 |
| COMMUNITY COORDINATOR | ACTIVE | 75010 | per Annum | 1820 | 72375.50 |
| 6 rows 9-17 of 17 columns | | | | | |

4. Converting data type in a columns

Before Cleaning:

Converting data types is another common task in data cleaning technique. This involves changing the data type of values in a column to a different type, such as from numeric to character or from character to numeric or factor and so on. Converting data types in a column can be a vital step in the data cleaning process. It can help ensure that your data is in a usable format, improve the performance of our analysis, avoid errors, and improve the appearance and readability of your data. After removing the special symbol from the BaseSalary, we checked that the BaseSalary column is showing as a character data type. We used the str () function to

check all the data types of the columns. We need the BaseSalary column to analyze the visualization. We want to change the data type of BaseSalary from a character to a numeric variable to analysis the data. The type of function is an alternative function used for checking the data type of a particular column.

Code:

```
#Converting character into num

#Check the data type of 'BaseSalary' using str function

str(payroll)
```

```
> str(payroll)
```

Screenshot:

```
> str(payroll)
tibble [205,735 × 17] (S3: tbl_df/tbl/data.frame)
 $ FiscalYear      : num [1:205735] 2020 2020 2020 2020 2020 2020 2020 2020 2020 2020 2020 2020
 $ PayrollNumber   : num [1:205735] 67 67 67 67 67 67 67 67 67 67 67 ...
 $ AgencyName      : chr [1:205735] "ADMIN FOR CHILDREN'S SVCS" "ADMIN FOR CHILDREN'S
 $ LastName        : chr [1:205735] "SEALEY" "RODRIGUEZ" "LEGAKIS" "ISOLA" ...
 $ FirstName       : chr [1:205735] "RHONDA" "EDWIN" "LUZ" "MICHAEL" ...
 $ MidInit         : chr [1:205735] "p" "j" "m" "j" ...
 $ AgencyStartDate : POSIXct[1:205735], format: "2015-06-01" "2018-12-10" "2013-10-28"
 $ WorkLocationBorough : chr [1:205735] "MANHATTAN" "MANHATTAN" "MANHATTAN" "MANHATTAN" ..
 $ TitleDescription : chr [1:205735] "ADMINISTRATIVE DIRECTOR OF SOCIAL SERVICES" "COMM
 $ LeaveStatusasofJune30: chr [1:205735] "ACTIVE" "ACTIVE" "CEASED" "ACTIVE" ...
 $ BaseSalary      : chr [1:205735] "97,755" "71,360" "65562" "140899" ...
 $ PayBasis        : chr [1:205735] "per Annum" "per Annum" "per Annum" "per Annum" ..
 $ RegularHours    : num [1:205735] 1820 1820 70 1820 1820 1820 1820 1820 1820 1820 ..
 $ RegularGrossPaid : num [1:205735] 96254 70265 2512 128671 109585 ...
 $ OTHours         : num [1:205735] 0 0 0 0 0 0 0 0 0 0 ...
 $ TotalOTPaid     : num [1:205735] 0 0 0 0 0 0 0 0 0 0 ...
 $ TotalOtherPay    : num [1:205735] 0 0 0 0 0 0 0 0 0 0 ...
 - attr(*, "na.action")= 'omit' Named int [1:392104] 1 2 3 6 15 34 35 37 39 40 ...
 ..- attr(*, "names")= chr [1:392104] "1" "2" "3" "6" ...
```

After Cleaning:

We used to convert the BaseSalary column data type using 'as.double' function from character to numeric data. The 'as.double' function is used to convert the data type of a column to double precision numeric values. It shows numbers with decimal places and allows for greater precision. Converting a BaseSalary column to double precision numeric values is often a necessary step in data cleaning because we used this column for statistical section which required that input data be in a specific format, such as double precision numeric values. By using the as.double function, we can ensure that our data is in the correct format and can be used in these statistical section without encountering errors.

Code:

```
#Converting character into num  
payroll$BaseSalary <- as.double(payroll$BaseSalary)  
str(payroll)  
head(payroll)
```

```
> payroll$BaseSalary <- as.double(payroll$BaseSalary)  
> str(payroll)
```

Screenshot:

```
> str(payroll)
tibble [240,562 × 17] (S3: tbl_df/tbl/data.frame)
 $ FiscalYear      : num [1:240562] 2020 2020 2020 2020 2020 2020 2020 2020 2020 2020 2020 2020
 $ PayrollNumber   : num [1:240562] 67 67 67 67 67 67 67 67 67 ...
 $ AgencyName      : chr [1:240562] "ADMIN FOR CHILDREN'S SVCS" "ADMIN FOR CHILDREN'S
 $ LastName        : chr [1:240562] "SEALEY" "SEALEY" "RODRIGUEZ" "LEGAKIS" ...
 $ FirstName       : chr [1:240562] "RHONDA" "RHONDA" "EDWIN" "LUZ" ...
 $ MidInit         : chr [1:240562] "P" "P" "J" "M" ...
 $ AgencyStartDate : Date[1:240562], format: "2015-06-01" "2015-06-01" "2018-12-10" "2
 $ WorkLocationBorough : chr [1:240562] "MANHATTAN" "MANHATTAN" "MANHATTAN" "MANHATTAN" ..
 $ TitleDescription : chr [1:240562] "ADMINISTRATIVE DIRECTOR OF SOCIAL SERVICES" "ADM
 $ LeaveStatusasofJune30: chr [1:240562] "ACTIVE" "ACTIVE" "ACTIVE" "CEASED" ...
 $ BaseSalary      : num [1:240562] 1 14.2 380.6 65562 140899 ...
 $ PayBasis        : chr [1:240562] "per Annum" "per Annum" "per Annum" "per Annum" ..
 $ RegularHours    : num [1:240562] 1820 1820 1820 70 1820 1820 1820 1820 1820 1820 1820 ..
 $ RegularGrossPaid : num [1:240562] 96254 96254 70265 2512 128671 ...
 $ OTHours         : num [1:240562] 0 0 0 0 0 0 0 0 0 0 ...
 $ TotalOTPaid     : num [1:240562] 0 0 0 0 0 0 0 0 0 0 ...
 $ TotalOtherPay   : num [1:240562] 0 0 0 0 0 0 0 0 0 0 ...
 attr(*, "na.action")= 

 Named int [1:240562] 1 2 3 6 15 24 25 27 29 40
```

```
> head(payroll)
# A tibble: 6 × 17
  FiscalYear PayrollNumber AgencyName      LastName FirstName MidInit AgencyStartDate workLocat... Title_2 Leave_3 BaseS_4 PayBa_5 Regul_6 Regul_7 O
    <dbl>      <dbl>      <chr>      <chr>      <chr>      <chr>      <date>      <chr>      <chr>      <chr>      <dbl> <chr>      <dbl> <dbl>
1    2020          67 ADMIN FOR CHILDREN'S SVCS SEALEY RHONDA P 2015-06-01 MANHATTAN ADMINI... ACTIVE 1.42e1 per An... 1820 96254.
2    2020          67 ADMIN FOR CHILDREN'S SVCS SEALEY RHONDA P 2015-06-01 MANHATTAN ADMINI... ACTIVE 1.42e1 per An... 1820 96254.
3    2020          67 ADMIN FOR CHILDREN'S SVCS RODRIGUEZ EDWIN J 2018-12-10 MANHATTAN COMMUN... ACTIVE 3.81e2 per An... 1820 70265.
4    2020          67 ADMIN FOR CHILDREN'S SVCS LEGAKIS LUZ M 2013-10-28 MANHATTAN COMMUN... CEASED 6.56e4 per An... 70 2512.
5    2020          67 ADMIN FOR CHILDREN'S SVCS ISOLA MICHAEL J 2011-04-25 MANHATTAN COMPUT... ACTIVE 1.41e5 per An... 1820 128671.
6    2020          67 ADMIN FOR CHILDREN'S SVCS MEDLEY LESLIE F 1996-06-23 MANHATTAN DIRECT... ACTIVE 1.14e5 per An... 1820 109585.
# ... with abbreviated variable names 'workLocationBorough', 'TitleDescription', 'LeaveStatusasofJune30', 'BaseSalary', 'PayBasis', 'RegularHours', 'RegularGrossPaid',
# 'TotalOtherPay'
```

ANALYSIS & VISUALIZATIONS

The importance of data analysis and visualization lies in their ability to help us extract insights and information from data, which can be used to make informed decisions and take appropriate actions. Data analysis involves using statistical, mathematical, and computational techniques to identify patterns, trends, and relationships in data, while data visualization involves using visual representations to communicate and represent the data in a way that is easy to understand and interpret. Together, data analysis and visualization are essential tools for working with data and for extracting valuable insights and information from data. We used Rstudio tool for analyzing dataset. Rstudio is a popular integrated development

environment (IDE) for the R programming language. It is widely used for data analysis and visualization because it provides several useful features and tools that make it easy to work with data in R. Moreover, Rstudio is commonly used for analyzing and visualizing data because it is free, user-friendly, customizable, and well-supported by a large community of users. We used various charts such as pie chart, bar charts, box plot chart which we have shown below to analyze the payroll dataset.

1. What is the percentage pay type basis in New York?

Screenshot:

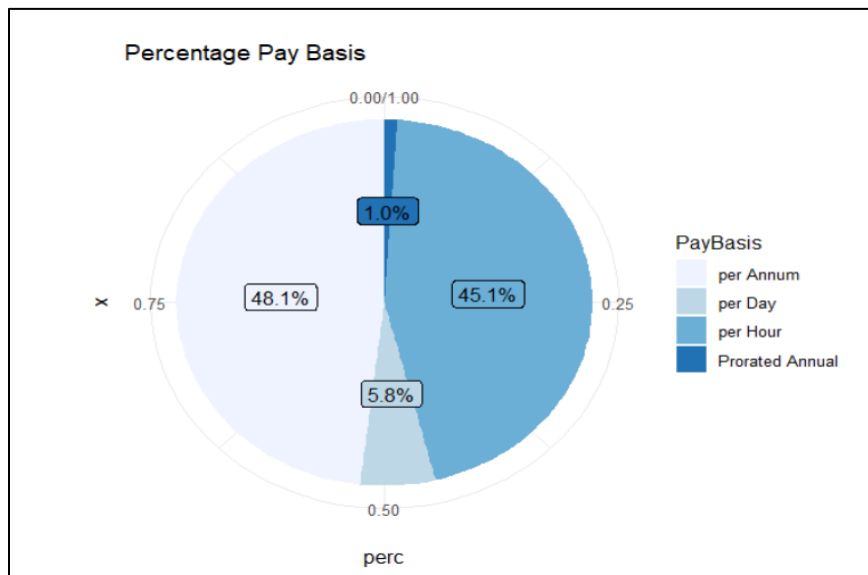


Figure 1 Pie Chart showing PayBasis type over the New York state

Insights:

The New York city Payroll contains information about the Pay Type basis. Which means the employees are paid on what basis are they paid per Hour, per Annum, per Day or Prorated

Annually. The above chart (Figure 1) shows the percentage of each Pay type basis. As per the visualization, it is evident that most of the employees are paid on per Annual basis which constitutes of almost half i.e., 48.1%. This number is almost equal when compared to per Hour Salaried employees i.e., 45.1%. This above visualization also shows that the percentage of employees paid on per Day basis is substantially low which is equivalent to 5 % and the Prorated Annually employees are the least, with only 1% of them. So, it can be concluded that in the city of New York most of the employees working in public sector are paid on per Annum or per Hour basis.

R Features Used:

- **Plot Type:** Pie Plot
- **Functions:** group_by, count, arrange, head, aesthetics(aes), coord_polar, geom_col, geom_label, labs, theme, scale_fill_brewer, theme_minimal
- **Libraries:** ggplot2, dplyr, scales, RColorBrewer

Code:

```
> df <- payroll %>% #Data Transformation
+ group_by(PayBasis) %>% # Variable to be transformed
+ count() %>%
+ ungroup() %>%
+ mutate(perc = `n` / sum(`n`)) %>%
+ arrange(perc) %>%
+ mutate (labels = scales::percent(perc))
> View(df)
```

| | PayBasis | n | perc | labels |
|---|-----------------|--------|-------------|--------|
| 1 | Prorated Annual | 2363 | 0.009822832 | 1.0% |
| 2 | per Day | 14038 | 0.058355019 | 5.8% |
| 3 | per Hour | 108398 | 0.450603171 | 45.1% |
| 4 | per Annum | 115763 | 0.481218979 | 48.1% |

PIE CHART

```
> library(ggplot2)

> ggplot(df, aes(x = "", y = perc, fill = PayBasis)) +
+   geom_col() +
+   geom_label(aes(label = labels),
+               position = position_stack(vjust = 0.5),
+               show.legend = FALSE) +
+   coord_polar(theta = "y") + labs(title = "Percentage Pay
Basis") + scale_fill_brewer(palette = "Blues") + theme_minimal()
```

```
> library(ggplot2)
>
> #Data Transformation
> df <- payroll %>%
+   group_by(PayBasis) %>% # Variable to be transformed
+   count() %>%
+   ungroup() %>%
+   mutate(perc = `n` / sum(`n`)) %>%
+   arrange(perc) %>%
+   mutate(labels = scales::percent(perc))
>
> View(df)
>
> # PIE CHART
>
> ggplot(df, aes(x = "", y = perc, fill = PayBasis)) +
+   geom_col() +
+   geom_label(aes(label = labels),
+               position = position_stack(vjust = 0.5),
+               show.legend = FALSE) +
+   coord_polar(theta = "y") + labs(title = "Percentage Pay Basis") + scale_fill_brewer(palette = "Blues") + theme_minimal()
```


2. Which are the top 5 work locations in terms of average regular gross paid?

Screenshot:

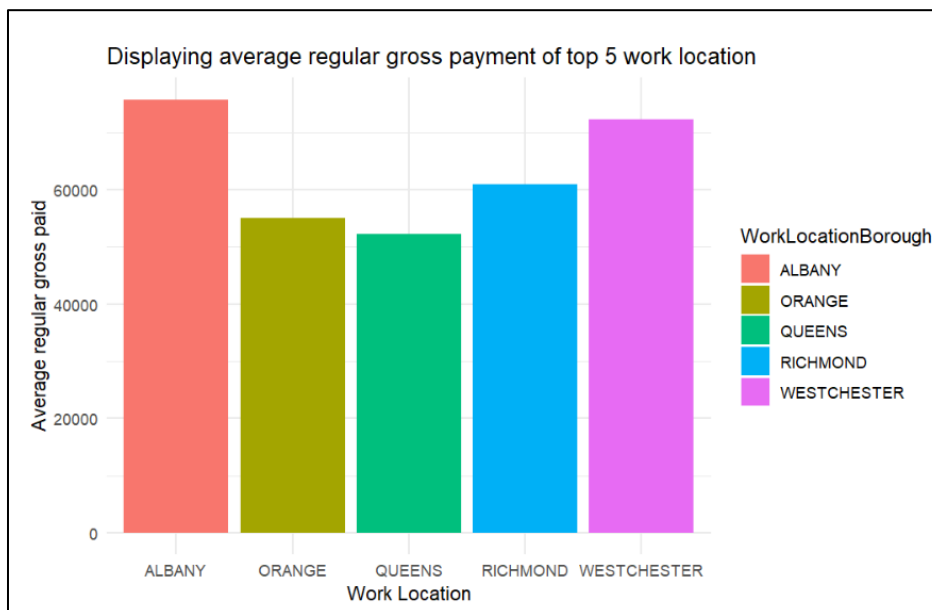


Figure 2 Vertical Bar Chart showing top 5 work location over the average regular gross payment

Insights:

Regular gross paid represents actual base salary during reporting period, which is the portion of the person's annual salary paid before deductions are calculated withheld. This does not include overtime pay or other compensation and does not reflect the after-tax amount or net pay. It is the amount paid to the employee for base salary during the fiscal year. According to the Visualization above (Figure 2), the Vertical Bar Graph Shows the top 5 work locations which has the highest Average Regular Gross Pay. Work Locations such as Albany, Orange, Queens, Richmond and Westchester have the highest average Regular Gross Pay. Where, Albany topping the list of Top 5 Locations with around \$75000 per annum and Queens acquires the least position among the Average Regular Gross Pay as \$52000 per annum.

R features:

- **Plot Type:** Bar Plot (Vertical)
- **Functions:** group_by, summarize, arrange, head, aesthetics(aes), fill, stat, labs, theme, geom_bar
- **Libraries:** ggplot2

Code:

```
> library(ggplot2)

>

> payroll %>%

+   group_by(WorkLocationBorough) %>%

+   summarize(AvgRegularGrossPaid=(mean(RegularGrossPaid))) %>%

+   arrange(desc(AvgRegularGrossPaid)) %>%

+   head(5) %>%

+ggplot(aes(x=WorkLocationBorough,y=AvgRegularGrossPaid,fill=WorkLocationBorough)

+       )+geom_bar(stat="identity")+ labs(title = "Displaying average regular gross payment of

+       top 5 work location ", x = "Work Location", y = "Average regular gross paid")+

+       theme_minimal()
```

```
> payroll %>%
+   group_by(WorkLocationBorough) %>%
+   summarize(AvgRegularGrossPaid=(mean(RegularGrossPaid))) %>%
+   arrange(desc(AvgRegularGrossPaid)) %>%
+   head(5) %>%
+   ggplot(aes(x=WorkLocationBorough,y=AvgRegularGrossPaid,fill=WorkLocationBorough))+geom_bar(stat
="identity")+ labs(title = "Displaying average regular gross payment of top 5 work location ", x =
"Work Location", y = "Average regular gross paid")+ theme_minimal()
```

```

> payroll %>%
+   group_by(WorkLocationBorough) %>%
+   summarize(AvgRegularGrossPaid=(mean(RegularGrossPaid))) %>%
+   arrange(desc(AvgRegularGrossPaid)) %>%
+   head(5) %>%
+   ggplot(aes(x=WorkLocationBorough,y=AvgRegularGrossPaid,fill=WorkLocationBorough))+geom_bar(stat="identity")

```

3. What is total regular gross pay with respect to different leave status?

Screenshot:

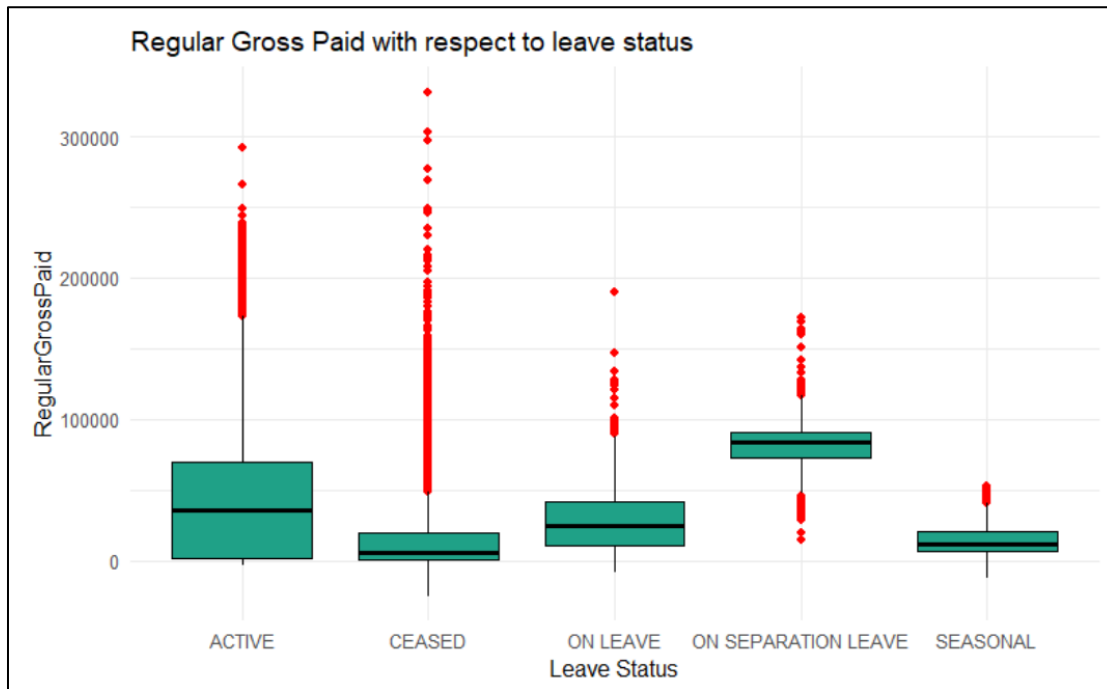


Figure 3 Boxplot Chart showing the distribution of regular gross paid with leave status

Insights:

The above Boxplot represents the Regular Gross Paid based on Leave Status of an employee. As it is evident from the graph that median Gross Pay of the employee who is terminated is lower than the other Leave Status Category. We can see that in the Ceased category the finish was much more spread out when compared to others. The Active Employees have the highest Gross Pay when compared to that of the Ceased Employee which has the lowest gross pay. We can also see from the graph that the median Gross pay for an employee who is on Seasonal leave is much

higher than that of an employee who is on leave or Seasonal. The box plot for 'On separation Leave', 'seasonal' and 'Ceased' is much shorter than the others which means that the range of lowest and the highest salary paid to employees of each of these categories. The box plot for employees with 'Active' leave status is comparatively tall, which suggests that the range of Lowest and highest Gross pay is the maximum in this case.

R features:

- **Plot Type:** Box & Whisker plot
- **Functions:** group_by, color, fill, aesthetics(aes), fill, outlier.color, labs, theme, geom_box
- **Libraries:** ggplot2, knitr

Code:

```
> library(knitr)
> library(ggplot2)
> #I generally apply the more aggressive solution by telling R to avoid all scientific
notation by setting options(scipen=999) at the top of the script. This forces the full display.
>
> options(scipen=999)
> format(summary payroll$RegularGrossPaid), big.mark = ",")
> payroll %>%
+ group_by(LeaveStatusasofJune30) %>%
+ ggplot(aes(x=LeaveStatusasofJune30,y= RegularGrossPaid))+geom_boxplot(color =
"black", fill = "#1FA187",outlier.colour = "red") +
```

```
+ labs(title = "Regular Gross Paid with respect to leave status",
+       x = "Leave Status") +
+ theme_minimal()
```

```
> options(scipen=999)
> format(summary payroll$RegularGrossPaid, big.mark = ",")
      Min.   1st Qu.   Median     Mean   3rd Qu.    Max.
"-25,104" "  1,575" " 28,696" " 37,140" " 64,118" "330,941"
>
> payroll %>%
+ group_by(LeaveStatusasofJune30) %>%
+ ggplot(aes(x=LeaveStatusasofJune30,y= RegularGrossPaid))+geom_boxplot(color = "black", fill = "#1FA187",outlier.colour = "red")
+
+ labs(title = "Regular Gross Paid with respect to leave status",
+       x = "Leave Status") +
+ theme_minimal()
```

4. What is the count of employees for the given fiscal year as per the PayBasis?

Screenshot:

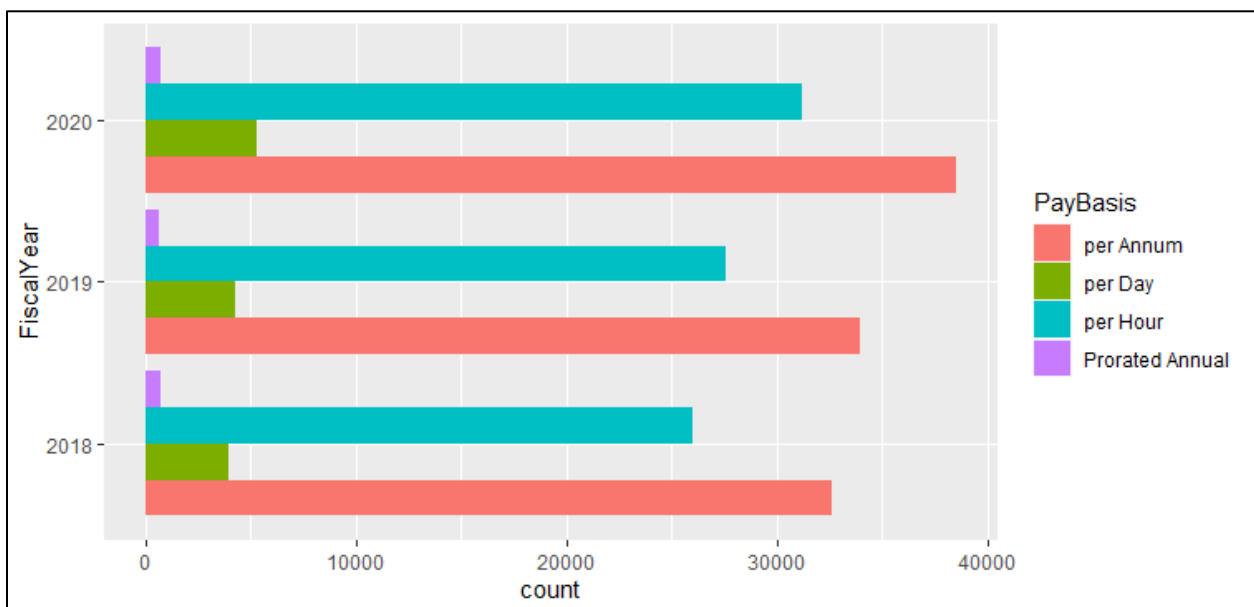


Figure 4 Horizontal Dodge Bar chart displaying count of employees for the given fiscal year as per the PayBasis

Insights:

The Figure 4 shows a Horizontal Dodge Bar Chart of Pay Basis per each Fiscal Year. We can conclude that the number of employees who are Prorated Annually remains almost the same across the three years. We can also see that the number of employees for all the other 'Pay Basis' Categories are increased every year. This also suggests that the total number of Employees in the public sector are gradually increasing with increasing year.

R features Used:

- **Plot Type:** Dodge Bar chart
- **Functions:** geom_bar(position="dodge"), coord_flip, fill
- **Libraries:** ggplot2, dplyr

Code:

```
> library(ggplot2)
> library(dplyr)
> p<-ggplot(payload,aes(FiscalYear, fill=PayBasis))
> p+geom_bar(position="dodge")+coord_flip()
```

```
> p<-ggplot(payload,aes(FiscalYear,fill=PayBasis))
> p+geom_bar(position="dodge")+coord_flip()
> |
```

STATISTICAL SUMMARY, SCRIPT, FUNCTIONS

1. Summary Statistical Functions

```
summary payroll)
```

```
> summary payroll)
  FiscalYear  PayrollNumber  AgencyName  LastName  FirstName  MidInit  AgencyStartDate  WorkLocationBorough  TitleDescription
Min. :2018    Min. : 3.0    Length:240562  Length:240562  Length:240562  Length:240562  Min. :1957-07-16  Length:240562  Length:240562
1st Qu.:2018   1st Qu.:300.0    Class :character  Class :character  Class :character  Class :character  1st Qu.:2006-11-20  Class :character  Class :character
Median :2019   Median :468.0    Mode :character  Mode :character  Mode :character  Mode :character  Median :2013-01-01  Mode :character  Mode :character
Mean :2019     Mean :506.7
3rd Qu.:2020   3rd Qu.:781.0
Max. :2020     Max. :902.0
LeaveStatusasofJune30  BaseSalary  PayBasis  RegularHours  RegularGrossPaid  OTHours  TotalOTPaid  TotalOtherPay
Length:240562  Min. : 1.0    Length:240562  Min. : -378.0  Min. : -25104  Min. : 0.00  Min. : 0  Min. : 0
Class :character  1st Qu.: 14.2  Class :character  1st Qu.: 0.0  1st Qu.: 1575  1st Qu.: 0.00  1st Qu.: 0  1st Qu.: 0
Mode :character  Median : 380.6  Mode :character  Median : 555.5  Median : 28696  Median : 0.00  Median : 0  Median : 0
Mean : 34739.9  Mean : 904.0  Mean : 37140  Mean : 80.39  Mean : 4349  Mean : 2775
3rd Qu.: 66388.0  3rd Qu.:1820.0  3rd Qu.: 64118  3rd Qu.: 55.00  3rd Qu.: 2344  3rd Qu.: 3307
Max. :293000.0  Max. :2936.0  Max. :330941  Max. :3147.00  Max. :189638  Max. :214703
```

Base Salary:

```
# Statistical Summary of Base Salary
summary payroll$`BaseSalary`)
min payroll$`BaseSalary`)
max payroll$`BaseSalary`)
mean payroll$`BaseSalary`)
median payroll$`BaseSalary`)
sd payroll$`BaseSalary`)
```

```
> summary payroll$`BaseSalary`)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.0   14.2   380.6 34739.9 66388.0 293000.0
> min payroll$`BaseSalary`)
[1] 1
> max payroll$`BaseSalary`)
[1] 293000
> mean payroll$`BaseSalary`)
[1] 34739.93
> median payroll$`BaseSalary`)
[1] 380.64
> sd payroll$`BaseSalary`)
[1] 40227.63
```

- **Mean:** Mean is the average of the given numbers and is calculated by dividing the sum of given numbers by the total number of numbers. $\text{Mean} = (\text{Sum of all the observations} / \text{Total number of observations})$. The mean value of the 'Base Salary' is \$34739.9.
- **Median:** The median is the value separating the higher half from the lower half of a data sample. The Median for 'Base Salary' is \$380.64.
- **Maximum:** The maximum value of a function is the place where a function reaches its highest point, or vertex, on a graph. and the Maximum 'Base Salary' of the Listing is \$293000
- **Minimum:** The minimum value is the smallest in the dataset. Of all the 'Base Salary' of the Listing provided in the dataset, \$1.00 is the least.
- **Standard Deviation:** A Standard Deviation is a measure of how dispersed the data is in relation to the mean. Standard deviation is important because it helps in understanding the measurements when the data is distributed. The more the data is distributed, the greater will be the standard deviation of that data. The Standard Deviation of the 'Base Salary' of all the listings is \$40277.63.

OT Hours:

```
#Statistical Summary for OT Hours
summary payroll$`OTHours`
min(payroll$`OTHours`)
max(payroll$`OTHours`)
mean(payroll$`OTHours`)
median(payroll$`OTHours`)
sd(payroll$`OTHours`)
```



```

> summary payroll$`OTHours`
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.00   0.00   0.00  80.39  55.00 3147.00
> min(payroll$`OTHours`)
[1] 0
> max(payroll$`OTHours`)
[1] 3147
> mean(payroll$`OTHours`)
[1] 80.39074
> median(payroll$`OTHours`)
[1] 0
> sd(payroll$`OTHours`)
[1] 174.7173

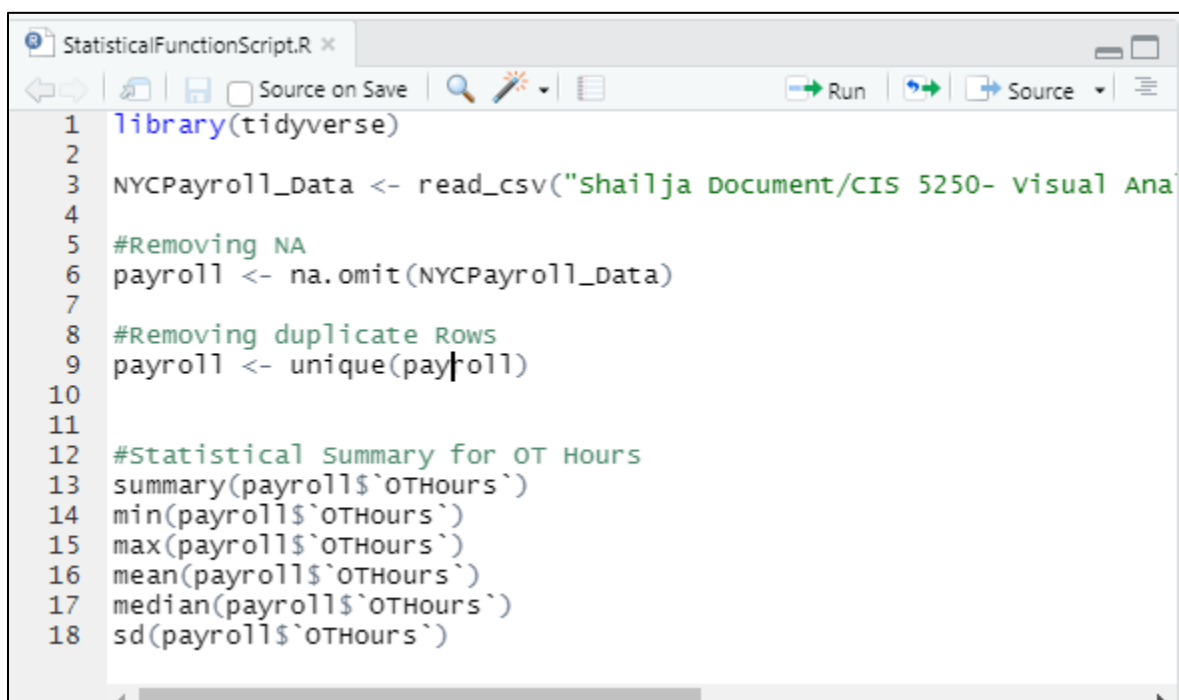
```

- **Mean:** Mean is the average of the given numbers and is calculated by dividing the sum of given numbers by the total number of numbers. $\text{Mean} = (\text{Sum of all the observations} / \text{Total number of observations})$. The mean value of the Listing 'OT Hours' is \$80.39074.
- **Median:** The median is the value separating the higher half from the lower half of a data sample. The Median for 'OT Hours' is 0.
- **Maximum:** The maximum value of a function is the place where a function reaches its highest point, or vertex, on a graph. and the Maximum 'OT Hours' of the Listing is \$3147.00
- **Minimum:** The minimum value is the smallest in the dataset. Of all the 'OT Hours' of the Listing provided in the dataset, \$0.00 is the least.
- **Standard Deviation:** A Standard Deviation is a measure of how dispersed the data is in relation to the mean. Standard deviation is important because it helps in understanding the measurements when the data is distributed. The more the data is distributed, the

greater will be the standard deviation of that data. The Standard Deviation of the 'OT Hours' of all the listings is \$174.7173.

2. R Script

An R script is simply a text file containing (almost) the same commands that you would enter on the command line of R. A script is simply a text file containing a set of commands and comments.



```
1 library(tidyverse)
2
3 NYCPayroll_Data <- read_csv("Shailja Document/CIS 5250- Visual Ana
4
5 #Removing NA
6 payroll <- na.omit(NYCPayroll_Data)
7
8 #Removing duplicate Rows
9 payroll <- unique(payroll)
10
11
12 #Statistical Summary for OT Hours
13 summary(payroll$`OTHours`)
14 min(payroll$`OTHours`)
15 max(payroll$`OTHours`)
16 mean(payroll$`OTHours`)
17 median(payroll$`OTHours`)
18 sd(payroll$`OTHours`)
```

The above screenshot is a R Script written specifically for getting Statistical Summary for the OT(Overtime) Hours.

3. User defined function

A User Defined function is a block of code that performs a particular task. R language allows a user to define function according to their use and call it whenever they need with no objection to number of times they can be called. R also has some inbuilt functions such as mean (), geom_bar () etc.,

Below is the code for a function ‘StackedChartFunction’. This function has seven parameters, and the user can call the function whenever they want to create a Stacked Bar Chart. Below are 2 examples in which we have created two different stacked bar charts using the Function ‘StackedChartFunction’ by merely passing the parameters.

#Function Definition

```
StackedChartFunction<-
```

```
function(df,xaxis,ColumnName,titleName,xaxisName,yaxisName,FillaxisName){  
  ggplot(df,aes(xaxis,fill=ColumnName))+geom_bar()+labs(title=titleName,x=xaxis  
  Name,y=yaxisName,fill=FillaxisName)  
}
```

Function Call 1

```
StackedChartFunction (payroll, payroll$FiscalYear,  
  payroll$WorkLocationBorough,"Count of employees per Work Location and  
  Fiscal Year","FiscalYear","Count of Employees","Work Location")
```

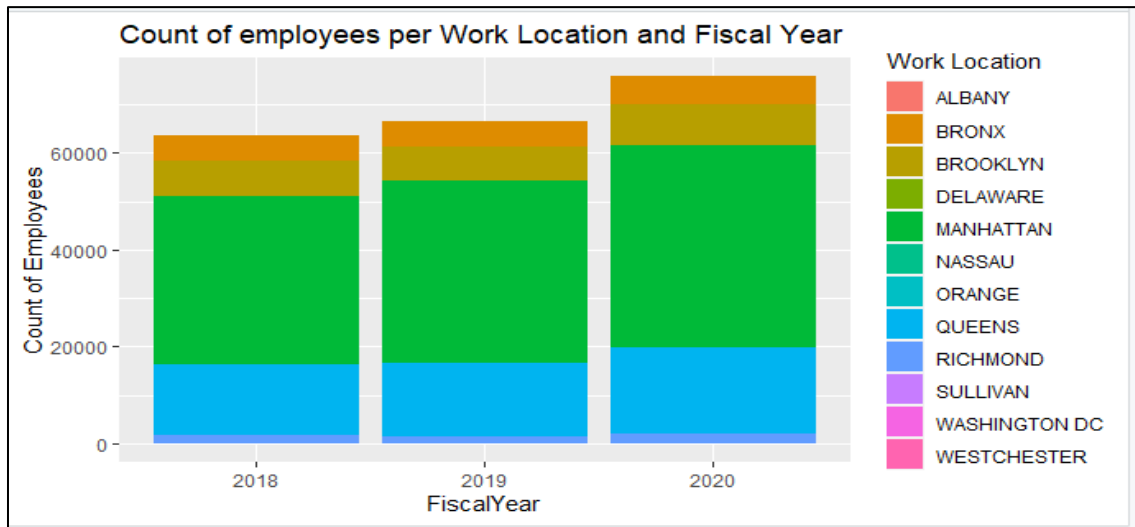
#Function Call 2

```
StackedChartFunction (payroll,payroll$FiscalYear,payroll$PayBasis,"Count of  
  employees per PayBasis and Fiscal Year","FiscalYear","Count of  
  Employees","Pay Basis")
```

Function Call 1

As it is seen in the code below, we have made a function call to the function ‘StackedChartFunction’ on the basis of Fiscal Year and Work Location.

```
> # Function Call 1
> StackedChartFunction payroll, payroll$FiscalYear, payroll$workLocationBorough, "Count of employees per work Location and Fiscal Year", "FiscalYear", "Count of Employees", "work Location")
>
```

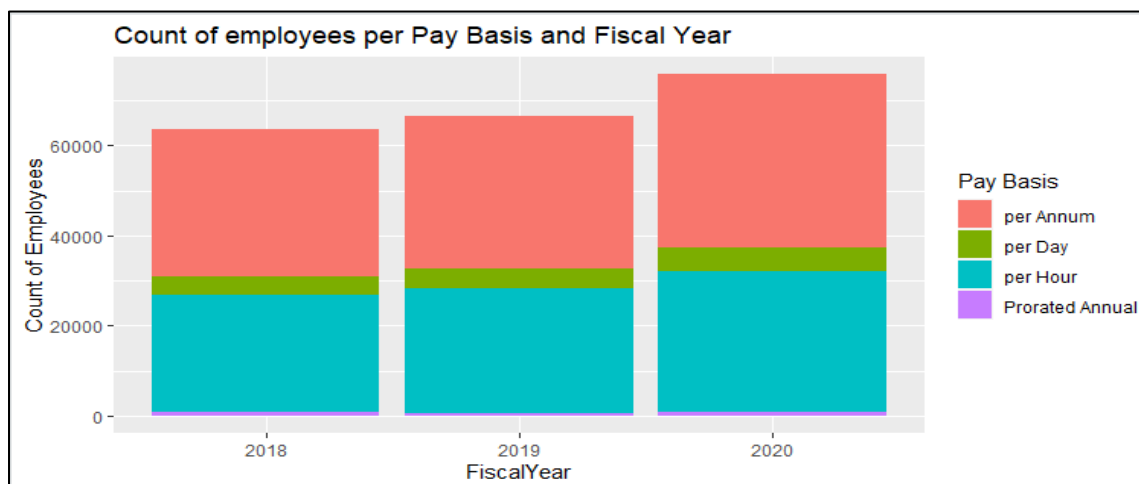


Function Call 2

As it is seen in the code below, we have made a function call to the function

‘StackedChartFunction’ on the basis of Fiscal Year and Pay Basis.

```
> # Function Call 2
> StackedChartFunction payroll, payroll$FiscalYear, payroll$PayBasis, "Count of employees per Pay Basis and Fiscal Year", "FiscalYear", "Count of Employees", "Pay Basis")
>
```



CONCLUSION

In this report we have analyzed the New York City Payroll data using R which is a statistical programming tool. We started with the motivation behind the analysis in the introduction then we found a publicly available dataset for the analysis. We then explore the data using different programming libraries and functions of R. We found that the data need lots of curation before we can start the analysis. We applied several data cleaning techniques like removing the Null/NA values, de-duplication of rows, fixed illegal entries like symbols and fixed the data types of various columns. After that, we started our analysis and visualization process. We used R studio along with various R libraries to visualize different types of charts like pie chart, bar chart and box plot chart. We visualized the percentage pay type basis in New York, the top 5 work locations in terms of average regular gross paid, the total regular gross pay with respect to different leave status and the top 10 work location based on the average base salary. Finally, we provided statistical summary for the payroll data, R script of the analysis and user defined function to with various function calls. This report provides the comprehensive analysis of the important attributes of the NYC Payroll data.

REFERNCES

1. Eliza Sayon (September 2021). 5 Advantages of working in the public sector
<https://publicspectrum.co/5-advantages-of-working-in-the-public-sector/>
2. Brock, J. (2001). United States Public Sector Employment. In: Dell’Aringa, C., Della Rocca, G., Keller, B. (eds) Strategic Choices in Reforming Public Service Employment. Palgrave Macmillan, London. https://doi.org/10.1057/9781403920171_5