

Employee's Payroll in Los Angeles

Team Members

Ankita Savaliya(CIN: 402659917)

Bhumika Suvagia(CIN: 402659969)

MSIS, California State University, Los Angeles

5250: Visual Analytics, Fall 2022

Professor Dr. Shilpa Balan

December 15, 2022

Table of Content

A. Introduction	2
B. Dataset URL	3
C. Data Description	4
D. Data Cleaning	7
E. Analysis & Visualizations	12
Figure 1	12
Figure 2	15
Figure 3	22
F. Statistical Summary, Script, Function	27
References	35

A. Introduction:

California is the first state in the United States to have recently passed a law that requires companies to post salary information, median gender, and racial pay gaps in all job postings, which inspired us to work deeply with Employees' Payroll in Los Angeles from the Kaggle website. As we know that Los Angeles is the second-largest metropolitan region in the United States, it would be interesting to know about wages and benefits in different jobs. Apart from that, it would also be intriguing to know how benefits and salaries differ for employees across departments and titles. We will also have the ability to find the number of employees by ethnicity and pay Distribution around it. In addition to that, we can determine full-time and part-time wage and hour efficiency based on regular and overtime pay distribution. This dataset is very helpful to get a better understanding of jobs in LA. We tried to analyze the aspects of this payroll using the following questions:

- What is the distribution of employees by gender in departments?
- What would be the distribution of pay in departments?
- Which ethnicity has the highest share of employment in all the departments?

B. Dataset URL:

The data set that we are going to use for the analysis is “payroll.csv” that we have loaded from Kaggle.com.

Dataset URL: <https://www.kaggle.com/dsfelix/employees-payroll-in-los-angeles/>

Data dictionary file: payroll.csv , **Size:** Row: 685462 and columns: 16

We imported our data set by setting a working directory as below

```
> setwd("~/Cal_State/5250/R Project/Dataset")
> LA_Payroll_data<-read.csv("LA_Payroll.csv")
> View(LA_Payroll_data)
```

```
> setwd("~/Cal_State/5250/R Project/Dataset")

> LA_Payroll_data<-read.csv("LA_Payroll.csv")

> View(LA_Payroll_data)
```

The below screenshot shows the whole data set after importing the CSV file in R Studio.

	RECORD_NBR	PAY_YEAR	DEPARTMENT_NO	DEPARTMENT_TITLE	JOB_CLASS_PGRADE	JOB_TITLE	EMPLOYMENT_TYPE	JOB_STATUS	MOU
1	3.0303E+11	2021	98	WATER AND POWER	7525-5	ELTL ENGR ASSOC	FULL_TIME	ACTIVE	3
2	3.03031E+11	2020	98	WATER AND POWER	1230-2	COML SRVC REPTV	FULL_TIME	ACTIVE	7
3	3036373631	2021	70	POLICE	2214-2	POLICE OFFICER II	FULL_TIME	ACTIVE	24
4	3.03138E+11	2020	98	WATER AND POWER	7207-3	SR CVL ENGG DRFTG TCHN	FULL_TIME	ACTIVE	2
5	3.03231E+11	2020	98	WATER AND POWER	1611-2	MTR RDR	FULL_TIME	ACTIVE	8
6	3135313039	2021	70	POLICE	2223-2	POLICE DETECTIVE II	FULL_TIME	ACTIVE	24
7	3.13632E+11	2021	88	RECREATION AND PARKS	3774-0	AIR CONDITIONING MECHANIC	FULL_TIME	ACTIVE	2
8	3136343834	2021	88	RECREATION AND PARKS	3141-0	GARDENER CARETAKER	FULL_TIME	ACTIVE	4
9	3.13635E+11	2021	70	POLICE	2207-1	POLICE SERVICE REPRESENTATIVE I	FULL_TIME	ACTIVE	3

MOU_TITLE	REGULAR_PAY	OVERTIME_PAY	ALL_OTHER_PAY	TOTAL_PAY	CITY_RETIREMENT_CONTRIBUTIONS	BENEFIT_PAY	GENDER	ETHNICITY
PROFESSIONAL UNIT	92210.52	23993.69	12596.92	128801.13	NA	6760.77	MALE	CAUCASIAN
CLERICAL UNIT	66291.60	41736.73	3918.57	111946.90	NA	24845.07	MALE	HISPANIC
POLICE OFFICERS, LIEUTENANT AND BELOW	133337.60	36874.05	1525.00	171736.65	62468.67	19749.08	MALE	BLACK
TECHNICAL REPRESENTATION UNIT	60557.69	6242.67	2182.48	68982.84	NA	18480.54	MALE	HISPANIC
OPERATING MAINTENANCE AND SERVICE UNIT	49245.36	12196.36	880.26	62321.98	NA	6443.85	MALE	CAUCASIAN
POLICE OFFICERS, LIEUTENANT AND BELOW	163176.00	31026.95	6545.00	200747.95	76447.96	19748.08	MALE	BLACK
BUILDING TRADES	98338.48	179.36	1494.22	100012.06	29167.19	19970.46	MALE	HISPANIC
EQUIPMENT OPERATION AND LABOR	60405.28	308.49	232.00	60945.77	17916.21	19178.16	MALE	CAUCASIAN
CLERICAL	57683.93	4510.77	2737.24	64931.94	1719.50	20112.00	FEMALE	BLACK

C. Data Description:

This Dataset contains complete Los Angeles Payroll information about 685,000 rows with regular, overtime, and all other pay (bonus, adjustments, and lump sum payouts) from 2013 to 2020. It also provides detailed information on benefits pay that includes health care, dental care, vision care, and life insurance. Moreover, this data includes some useful categories such as department title, employment type, gender, ethnicity, etc. Apart from this, out of 18 columns, 16 columns are being used in this R Project.

The below table describes the data set in detail.

Data Field	Description	Example Values
RECORD_NBR	A unique number to identify an employee Data type: Integer	303030000000, 3030313337
PAY YEAR	The tax-year employee was paid Data type: Integer	2020,2021
DEPARTMENT_NO	Department Number in the city payroll system Data type: Integer	Starting from 2 to 98
DEPARTMENT_TITLE	Title of city department Data type: Character	Water And Power, Aging
JOB_TITLE	Employee's job title Data type: Character	Police Officer II, Recreation Assistance

EMPLOYMENT_TYPE	Employee's employment type Data type: Character	Full Time, Part Time, Per Event
JOB_STATUS	Employee's job status at the time the data was uploaded Data type: Character	Active, Non-Active
MOU_TITLE	Employee's title of Memorandum of Understanding Data type: Character	Police Officers, Lieutenant and Below, Clerical
REGULAR_PAY	Employee's fixed payment Data type: Numeric	Approximately Up to \$4,00,000
OVERTIME_PAY	Payments attributable to hours worked beyond the regular work schedule Data type: Numeric	Between \$0 to \$434K
ALL_OTHER_PAY	Any payments are other than Regular and Overtime. This includes bonuses, adjustments, and lump sum payouts Data type: Numeric	Between \$0 to \$2.39M
TOTAL_PAY	Total payment includes Regular, Overtime, and all other payments	Approximately Up to \$2,00,000
CITY_RETIREMENT_C ONTRIBUTIONS	Estimated payments made by the city towards employee's retirement Data type: Numeric	Between \$0 to \$164K
BENEFIT_PAY	City contribution to the employee's health care, dental care, vision care,	Between \$0 to \$58.8K

	and life insurance Data type: Numeric	
GENDER	Gender as self-reported by an employee Data type: Character	Male, Female
ETHNICITY	Employee's ethnicity Data type: Character	White, Black, or African American, American Indian or Alaska Native, Asian

D. Data Cleaning:

After observing the raw data, we found out there are almost 9000 “NA” values, unnecessary columns in our dataset. In addition to this, we discovered certain fields had duplicate values, therefore we cleaned the dataset using the below steps in R studio.

1) Removing multiple columns:

Since our main aim was to analyze the different department payrolls and employees, we decided to remove the “JOB_CLASS_PGRADE” and “MOU” fields from the data set as they were not useful columns.

Before cleaning:

#	RECORD_NBR	PAY_YEAR	DEPARTMENT_NO	DEPARTMENT_TITLE	JOB_CLASS_PGRADE	JOB_TITLE	EMPLOYMENT_TYPE	JOB_STATUS	MOU	MOU_TITLE
1	3.0303E+11	2021	98	WATER AND POWER	7525-5	ELTL ENGR ASSOC	FULL_TIME	ACTIVE	7	PROFESSIONAL UNIT
2	3.03031E+11	2020	98	WATER AND POWER	1230-2	COML SRVC REPTV	FULL_TIME	ACTIVE	7	CLERICAL UNIT
3	3036373631	2021	70	POLICE	2214-2	POLICE OFFICER II	FULL_TIME	ACTIVE	24	POLICE OFFICERS, LIEUTENANT AND BELOW
4	3.03138E+11	2020	98	WATER AND POWER	7207-3	SR CVL ENGG DRFTG TCHN	FULL_TIME	ACTIVE	2	TECHNICAL REPRESENTATION UNIT
5	3.03231E+11	2020	98	WATER AND POWER	1611-2	MTR RDR	FULL_TIME	ACTIVE	8	OPERATING MAINTENANCE AND SERVICE UNIT
6	3135313039	2021	70	POLICE	2223-2	POLICE DETECTIVE II	FULL_TIME	ACTIVE	24	POLICE OFFICERS, LIEUTENANT AND BELOW
7	3.13632E+11	2021	88	RECREATION AND PARKS	3774-0	AIR CONDITIONING MECHANIC	FULL_TIME	ACTIVE	2	BUILDING TRADES
8	3136343834	2021	88	RECREATION AND PARKS	3141-0	GARDENER CARETAKER	FULL_TIME	ACTIVE	4	EQUIPMENT OPERATION AND LABOR
9	3.13635E+11	2021	70	POLICE	2207-1	POLICE SERVICE REPRESENTATIVE I	FULL_TIME	ACTIVE	7	CLERICAL
10	3.13635E+11	2021	70	POLICE	2214-2	POLICE OFFICER II	FULL_TIME	ACTIVE	24	POLICE OFFICERS, LIEUTENANT AND BELOW
11	CTNA961968	2020	98	WATER AND POWER	920-1	CONSTR EQPT OPR	FULL_TIME	NOT_ACTIVE	2	DAILY RATE
12	3138333032	2021	70	POLICE	2223-2	POLICE DETECTIVE II	FULL_TIME	ACTIVE	24	POLICE OFFICERS, LIEUTENANT AND BELOW
13	3.03336E+11	2020	98	WATER AND POWER	7248-5	WTRWKS ENGR	FULL_TIME	ACTIVE	6	SUPERVISORY PROFESSIONAL UNIT
14	3138343539	2021	70	POLICE	2214-2	POLICE OFFICER II	FULL_TIME	ACTIVE	24	POLICE OFFICERS, LIEUTENANT AND BELOW
15	3.03339E+11	2020	98	WATER AND POWER	3558-5	PWR SHVL OPR	FULL_TIME	ACTIVE	8	OPERATING MAINTENANCE AND SERVICE UNIT
16	3.03432E+11	2020	98	WATER AND POWER	5854-5	WTR UTILITY OPR	FULL_TIME	ACTIVE	8	STEAM PLANT AND WATER SUPPLY UNIT
17	3.03031E+11	2021	98	WATER AND POWER	3786-5	STM PLT MTNC SUPV	FULL_TIME	ACTIVE	6	SUPERVISORY BLUE COLLAR UNIT
18	3.03534E+11	2020	98	WATER AND POWER	4260-5	CHF SFTY ENGR PRSR VSL5	FULL_TIME	NOT_ACTIVE	2	TECHNICAL REPRESENTATION UNIT
19	3.03539E+11	2020	98	WATER AND POWER	7209-2	SR ELTL ENGG DRFTG TCHN	FULL_TIME	ACTIVE	2	TECHNICAL REPRESENTATION UNIT

To remove unnecessary columns from the data set, we selected columns that we wanted to remove by index.

```
> setwd("~/Cal_State/5250/R Project/Dataset")
> LA_Payroll<-read.csv("LA_Payroll.csv")
> View(LA_Payroll)
> LA_Payroll_data<-LA_Payroll[, -c(5,9)]
> View(LA_Payroll_data)
```



```
> setwd("~/Cal_State/5250/R Project/Dataset")

> LA_Payroll<-read.csv("LA_Payroll.csv")

> View(LA_Payroll)

> LA_Payroll_data<-LA_Payroll[, -c(5,9)]

> View(LA_Payroll_data)
```

After cleaning:

After dropping out-of-scope data, we got the below dataset.

#	RECORD_NBR	PAY_YEAR	DEPARTMENT_NO	DEPARTMENT_TITLE	JOB_TITLE	EMPLOYMENT_TYPE	JOB_STATUS	MOU_TITLE	REGULAR_PAY	OVERTIME_PAY
1	3.0303E+11	2021	98	WATER AND POWER	ELTL ENGR ASSOC	FULL_TIME	ACTIVE	PROFESSIONAL UNIT	92210.52	23993.69
2	3.03031E+11	2020	98	WATER AND POWER	COML SRVC REPTV	FULL_TIME	ACTIVE	CLERICAL UNIT	66291.60	41736.73
3	3036373631	2021	70	POLICE	POLICE OFFICER II	FULL_TIME	ACTIVE	POLICE OFFICERS, LIEUTENANT AND BELOW	133337.60	36874.05
4	3.03138E+11	2020	98	WATER AND POWER	SR CVL ENGG DRFTG TCHN	FULL_TIME	ACTIVE	TECHNICAL REPRESENTATION UNIT	60557.69	6242.67
5	3.03231E+11	2020	98	WATER AND POWER	MTR RDR	FULL_TIME	ACTIVE	OPERATING MAINTENANCE AND SERVICE UNIT	49245.36	12196.36
6	3135313039	2021	70	POLICE	POLICE DETECTIVE II	FULL_TIME	ACTIVE	POLICE OFFICERS, LIEUTENANT AND BELOW	163176.00	31026.95
7	3.13632E+11	2021	88	RECREATION AND PARKS	AIR CONDITIONING MECHANIC	FULL_TIME	ACTIVE	BUILDING TRADES	98338.48	179.36
8	3136343834	2021	88	RECREATION AND PARKS	GARDENER CARETAKER	FULL_TIME	ACTIVE	EQUIPMENT OPERATION AND LABOR	60405.28	308.49
9	3.13635E+11	2021	70	POLICE	POLICE SERVICE REPRESENTATIVE I	FULL_TIME	ACTIVE	CLERICAL	57683.93	4510.77
10	3.13635E+11	2021	70	POLICE	POLICE OFFICER II	FULL_TIME	ACTIVE	POLICE OFFICERS, LIEUTENANT AND BELOW	99956.80	3895.39
11	CTNA961968	2020	98	WATER AND POWER	CONSTR EQPT OPR	FULL_TIME	NOT_ACTIVE	DAILY RATE	37729.93	0.00
12	3138333032	2021	70	POLICE	POLICE DETECTIVE II	FULL_TIME	ACTIVE	POLICE OFFICERS, LIEUTENANT AND BELOW	142436.62	21314.28
13	3.03336E+11	2020	98	WATER AND POWER	WTRWKS ENGR	FULL_TIME	ACTIVE	SUPERVISORY PROFESSIONAL UNIT	101862.63	8021.43
14	3138343539	2021	70	POLICE	POLICE OFFICER II	FULL_TIME	ACTIVE	POLICE OFFICERS, LIEUTENANT AND BELOW	118361.19	1718.70
15	3.03339E+11	2020	98	WATER AND POWER	PWR SHVL OPR	FULL_TIME	ACTIVE	OPERATING MAINTENANCE AND SERVICE UNIT	73394.40	2043.36
16	3.03432E+11	2020	98	WATER AND POWER	WTR UTILITY OPR	FULL_TIME	ACTIVE	STEAM PLANT AND WATER SUPPLY UNIT	67794.25	42388.79
17	3.03031E+11	2021	98	WATER AND POWER	STM PLT MTNC SUPV	FULL_TIME	ACTIVE	SUPERVISORY BLUE COLLAR UNIT	93660.96	46173.69
18	3.03534E+11	2020	98	WATER AND POWER	CHF SFTY ENGR PRSR VSLS	FULL_TIME	NOT_ACTIVE	TECHNICAL REPRESENTATION UNIT	96827.88	18683.13
19	3.03539E+11	2020	98	WATER AND POWER	SR ELTL ENGG DRFTG TCHN	FULL_TIME	ACTIVE	TECHNICAL REPRESENTATION UNIT	66485.63	14037.99

2) Deleting duplicate values:

We noticed that there were some duplicate values in the data set, so in order to delete those duplicate data, we used the “unique” function.

Before cleaning:

	RECORD_NBR	PAY_YEAR	DEPARTMENT_NO	DEPARTMENT_TITLE	JOB_TITLE	EMPLOYMENT_TYPE	JOB_STATUS	MOU_TITLE	REGULAR_PAY	
	1169	3334353936	2021	70	POLICE	DETENTION OFFICER	FULL_TIME	ACTIVE	SAFETY / SECURITY	84546.10
	1170	3334353937	2021	70	POLICE	POLICE SERVICE REPRESENTATIVE II	FULL_TIME	ACTIVE	CLERICAL	87945.08
	1171	3.43135E+11	2020	98	WATER AND POWER	CVL ENGG ASSO	FULL_TIME	ACTIVE	PROFESSIONAL UNIT	75201.80
	1172	3334373731	2021	70	POLICE	POLICE SERGEANT I	FULL_TIME	ACTIVE	POLICE OFFICERS, LIEUTENANT AND BELOW	137303.35
	1173	3.43136E+11	2020	98	WATER AND POWER	INSTRMT MCHC	FULL_TIME	ACTIVE	STEAM PLANT AND WATER SUPPLY UNIT	79004.65
	1174	3334373939	2021	70	POLICE	POLICE OFFICER III	FULL_TIME	ACTIVE	POLICE OFFICERS, LIEUTENANT AND BELOW	127230.26
	1175	3.03435E+11	2021	70	POLICE	DETENTION OFFICER	FULL_TIME	ACTIVE	SAFETY / SECURITY	56742.30
	1176	3334373939	2021	70	POLICE	POLICE OFFICER III	FULL_TIME	ACTIVE	POLICE OFFICERS, LIEUTENANT AND BELOW	127230.26
	1177	3.03435E+11	2021	70	POLICE	DETENTION OFFICER	FULL_TIME	ACTIVE	SAFETY / SECURITY	56742.30
	1178	3034353937	2021	70	POLICE	SENIOR SYSTEMS ANALYST I	FULL_TIME	ACTIVE	SUPERVISORY ADMINISTRATIVE	95400.45
	1179	3.03436E+11	2021	70	POLICE	POLICE OFFICER II	FULL_TIME	ACTIVE	POLICE OFFICERS, LIEUTENANT AND BELOW	78171.84
	1180	3.33438E+11	2021	70	POLICE	POLICE OFFICER II	FULL_TIME	ACTIVE	POLICE OFFICERS, LIEUTENANT AND BELOW	81676.80
	1181	3334383731	2021	70	POLICE	POLICE LIEUTENANT II	FULL_TIME	NOT_ACTIVE	POLICE OFFICERS, LIEUTENANT AND BELOW	62238.25
	1182	3.43231E+11	2020	98	WATER AND POWER	SENIOR ADMINISTRATIVE CLERK	FULL_TIME	ACTIVE	CLERICAL UNIT	48813.60
	1183	3.43232E+11	2020	98	WATER AND POWER	CVL ENGG ASSO	FULL_TIME	NOT_ACTIVE	PROFESSIONAL UNIT	83478.40
	1184	3.3353E+11	2021	70	POLICE	POLICE OFFICER II	FULL_TIME	ACTIVE	POLICE OFFICERS, LIEUTENANT AND BELOW	98074.80
	1185	3432333435	2020	98	WATER AND POWER	UTILITY SRVCS SPECIALIST	FULL_TIME	ACTIVE	TECHNICAL REPRESENTATION UNIT	82185.50

The below code was performed in the console

```
> LA_Payroll<-unique(LA_Payroll)
> dim(LA_Payroll)
[1] 51911    16
> View(LA_Payroll)
```

```
> LA_Payroll<-unique(LA_Payroll)

> dim(LA_Payroll)

[1] 51911    16

> View(LA_Payroll)
```

	RECORD_NBR	PAY_YEAR	DEPARTMENT_NO	DEPARTMENT_TITLE	JOB_TITLE	EMPLOYMENT_TYPE	JOB_STATUS	MOU_TITLE	REGULAR_PAY
1169	3334353936	2021	70	POLICE	DETENTION OFFICER	FULL_TIME	ACTIVE	SAFETY / SECURITY	84546.10
1170	3334353937	2021	70	POLICE	POLICE SERVICE REPRESENTATIVE II	FULL_TIME	ACTIVE	CLERICAL	87945.08
1171	3.43135E+11	2020	98	WATER AND POWER	CVL ENGG ASSO	FULL_TIME	ACTIVE	PROFESSIONAL UNIT	75201.80
1172	3334373731	2021	70	POLICE	POLICE SERGEANT I	FULL_TIME	ACTIVE	POLICE OFFICERS, LIEUTENANT AND BELOW	137303.35
1173	3.43136E+11	2020	98	WATER AND POWER	INSTRMT MCHC	FULL_TIME	ACTIVE	STEAM PLANT AND WATER SUPPLY UNIT	79004.65
1174	3334373939	2021	70	POLICE	POLICE OFFICER III	FULL_TIME	ACTIVE	POLICE OFFICERS, LIEUTENANT AND BELOW	127230.26
1175	3.03435E+11	2021	70	POLICE	DETENTION OFFICER	FULL_TIME	ACTIVE	SAFETY / SECURITY	56742.30
1178	3034353937	2021	70	POLICE	SENIOR SYSTEMS ANALYST I	FULL_TIME	ACTIVE	SUPERVISORY ADMINISTRATIVE	95408.45
1179	3.03436E+11	2021	70	POLICE	POLICE OFFICER II	FULL_TIME	ACTIVE	POLICE OFFICERS, LIEUTENANT AND BELOW	78171.84
1180	3.33438E+11	2021	70	POLICE	POLICE OFFICER II	FULL_TIME	ACTIVE	POLICE OFFICERS, LIEUTENANT AND BELOW	81676.80
1181	3334383731	2021	70	POLICE	POLICE LIEUTENANT II	FULL_TIME	NOT_ACTIVE	POLICE OFFICERS, LIEUTENANT AND BELOW	62238.25
1182	3.43231E+11	2020	98	WATER AND POWER	SENIOR ADMINISTRATIVE CLERK	FULL_TIME	ACTIVE	CLERICAL UNIT	48813.60
1183	3.43232E+11	2020	98	WATER AND POWER	CVL ENGG ASSO	FULL_TIME	NOT_ACTIVE	PROFESSIONAL UNIT	83478.40
1184	3.3353E+11	2021	70	POLICE	POLICE OFFICER II	FULL_TIME	ACTIVE	POLICE OFFICERS, LIEUTENANT AND BELOW	98074.80
1185	3432333435	2020	98	WATER AND POWER	UTLTY SRVCS SPECIALIST	FULL_TIME	ACTIVE	TECHNICAL REPRESENTATION UNIT	82185.50
1186	3335313633	2021	70	POLICE	POLICE OFFICER III	FULL_TIME	ACTIVE	POLICE OFFICERS, LIEUTENANT AND BELOW	128060.00
1187	3.03436E+11	2021	70	POLICE	CRIMINALIST II	FULL_TIME	ACTIVE	PROFESSIONAL ENGINEERING AND SCIENTIFIC	130741.04

3) Replacing "NA" values with 0:

We noticed that the "CITY RETIREMENT CONTRIBUTION" column contained many "NA" values. To replace NA with 0 in the data set, we used the `is.na()` function, then we selected all those "NA" values and assigned them to 0.

Before Cleaning

OVERTIME_PAY	ALL_OTHER_PAY	TOTAL_PAY	CITY_RETIREMENT_CONTRIBUTIONS	BENEFIT_PAY	GENDER
23993.69	12596.92	128801.13	NA	6760.77	MALE
41736.73	3918.57	111946.90	NA	24845.07	MALE
36874.05	1525.00	171736.65	62468.67	19749.08	MALE
6242.67	2182.48	68982.84	NA	18480.54	MALE
12196.36	880.26	62321.98	NA	6443.85	MALE
31026.95	6545.00	200747.95	76447.96	19748.08	MALE
179.36	1494.22	100012.06	29167.19	19970.46	MALE

```
> LA_Payroll_data<-read.csv("LA_Payroll.csv")
> LA_Payroll_data$CITY_RETIREMENT_CONTRIBUTIONS
 [1] NA NA 62468.67 NA NA 76447.96 29167.19 17916.21 1719.50 46829.76 NA 66731.56 6723.00
 [22] 7162.00 766.00 4991.00 NA NA 64869.40 2543.10 65784.15 58725.91 631.58 NA 5871.00 23279.58
 [43] 56461.37 5125.00 17412.20 6382.82 6141.00 4784.70 4842.42 NA NA 5631.00 675.00 NA 4725.00
 [64] NA 648.00 4811.00 4247.00 541.00 34621.79 NA 5619.00 651.00 4742.00 481.24 NA 2457.66
 [85] 39681.10 5744.00 NA 481.24 NA 4526.00 63179.18 5785.38 437.00 47723.11 NA 56461.37 1719.76
 [106] 2286.13 3877.00 56647.12 59454.52 NA 39619.36 52786.80 5649.00 6917.00 19331.60 475.00 61536.16 6734.00
 [127] 25339.18 71289.26 423.00 76447.96 NA 7151.00 5913.12 21589.60 1555.00 5441.00 5787.20 8197.00 2953.60
 [148] 1792.53 5945.00 25186.56 37466.70 43965.89 62267.20 56461.37 5293.77 6463.00 7626.00 61536.16 NA 32141.85
 [169] 37237.94 9679.00 59955.77 18351.30 NA 28853.46 474.00 NA 2476.98 4629.17 631.58 952.00 49458.15
 [190] 2579.81 53322.59 34393.41 5575.00 NA 4527.00 39534.65 NA NA 66731.56 477.27 54163.59 51935.00
 [211] 447.00 55693.78 NA 7387.00 5967.38 734.00 53754.77 56163.12 16635.29 9167.00 4422.00 4284.00 5357.20
 [232] NA 34734.90 4964.14 846.00 33842.57 NA 6128.00 53671.36 NA 1131.00 7646.00 55935.90 2966.22
 [253] 58156.97 43668.88 72432.35 64139.98 5229.64 48622.94 66731.56 NA 4699.66 584.66 NA 43758.37 39.15
 [274] 24543.59 6278.26 2977.45 24148.46 64996.32 62272.15 3989.94 4159.34 59984.49 NA NA 57335.40 66731.56
 [295] 37856.47 24119.00 NA NA 46765.44 NA 15838.91 3837.00 18125.34 44689.47 17916.21 1646.95 NA
 [316] 29614.84 2471.57 9253.00 29999.78 66731.56 961.00 6272.80 38917.73 53868.85 59534.73 53292.60 54728.50 6664.00
 [337] 5266.15 63179.18 4149.00 9679.00 824.00 65467.74 773.98 38744.10 NA 22819.61 4884.79 NA 4565.00
```

```
> LA_Payroll_data$CITY_RETIREMENT_CONTRIBUTIONS
```

```
[1] NA NA 62468.67 NA NA 76447.96 29167.19 17916.21 1719.50 46829.76 NA
66731.56 6723.00 55452.22 NA 4474.00 6182.00 6391.00 NA 28672.70 4218.71
```

```
[22] 7162.00 766.00 4991.00 NA NA 64869.40 2543.10 65784.15 58725.91 631.58
NA 5871.00 23279.58 2384.64 16743.57 26735.95 4312.00 2449.75 64139.98 3656.96 NA
```

```
[43] 56461.37 5125.00 17412.20 6382.82 6141.00 4784.70 4842.42 NA NA 5631.00
675.00 NA 4725.00 18193.63 5329.84 NA 63179.18 3999.11 2423.29 55717.79 7679.00
```

```
[64] NA 648.00 4811.00 4247.00 541.00 34621.79 NA 5619.00 651.00 4742.00
481.24 NA 2457.66 17124.50 77371.46 NA 29747.88 NA 538.00 3451.00 6542.00
```

```

> LA_Payroll_data[is.na(LA_Payroll_data)]<-0
> LA_Payroll_data$CITY_RETIREMENT_CONTRIBUTIONS
[1] 0.00 0.00 62468.67 0.00 0.00 76447.96 29167.19 17916.21 1719.50 46829.76 0.00 66731.56
[30] 58725.91 631.58 0.00 5871.00 23279.58 2384.64 16743.57 26735.95 4312.00 2449.75 64139.98 3656.96
[59] 63179.18 3999.11 2423.29 55717.79 7679.00 0.00 648.00 4811.00 4247.00 541.00 34621.79 0.00
[88] 481.24 0.00 4526.00 63179.18 5785.38 437.00 47723.11 0.00 56461.37 1719.76 0.00 27291.44
[117] 61536.16 6734.00 0.00 5715.00 0.00 68621.39 57335.40 5521.00 5963.00 18364.49 25339.18 71289.26
[146] 3283.53 5277.00 1792.53 5945.00 25186.56 37466.70 43965.89 62267.20 56461.37 5293.77 6463.00 7626.00
[175] 474.00 0.00 2476.98 4629.17 631.58 952.00 49458.15 33821.35 63179.18 0.00 51813.10 65784.15
[204] 2263.70 0.00 6125.42 6159.00 64996.32 4188.00 3856.00 447.00 55693.78 0.00 7387.00 5967.38
[233] 34734.90 4964.14 846.00 33842.57 0.00 6128.00 53671.36 0.00 1131.00 7646.00 55935.90 2966.22
[262] 584.66 0.00 43758.37 39.15 61456.90 66731.56 36522.37 0.00 22817.82 533.28 7337.00 55127.46
[291] 445.00 8585.00 58849.59 57166.30 37856.47 24119.00 0.00 0.00 46765.44 0.00 15838.91 3837.00
[320] 66731.56 961.00 6272.80 38917.73 53868.85 59534.73 53292.60 54728.50 6664.00 4565.00 0.00 3271.00
[349] 4565.00 55731.26 59733.90 57722.57 8899.91 64712.97 26311.80 5828.13 1667.68 56116.12 5293.77 22356.92
[378] 8675.70 27886.85 2939.50 5855.20 4328.00 5721.00 53365.15 2967.46 64996.32 59678.65 67519.39 39788.23
[407] 72432.35 5566.00 0.00 658.00 68621.38 69185.83 54257.60 66731.56 72432.35 4421.66 17916.21 66731.56
[436] 527.00 0.00 5864.21 2772.91 32448.20 3278.00 5487.00 0.00 3898.00 0.00 0.00 53671.36
[465] 7576.00 4246.00 0.00 0.00 38184.43 5597.00 55318.70 66731.56 0.00 0.00 34842.80 49625.42
[494] 63179.18 52421.78 49759.44 0.00 0.00 55397.40 26115.86 0.00 17295.20 865.72 0.00 4817.62
[523] 64939.94 0.00 4148.00 0.00 38383.49 59535.56 0.00 5224.00 0.00 26447.44 0.00 0.00

```

```
> LA_Payroll_data[is.na(LA_Payroll_data)]<-0
```

```
> LA_Payroll_data$CITY_RETIREMENT_CONTRIBUTIONS
```

```

[1] 0.00 0.00 62468.67 0.00 0.00 76447.96 29167.19 17916.21 1719.50
46829.76 0.00 66731.56 6723.00 55452.22 0.00 4474.00 6182.00 6391.00 0.00
28672.70 4218.71 7162.00 766.00 4991.00 0.00 0.00 64869.40 2543.10 65784.15

```

```

[30] 58725.91 631.58 0.00 5871.00 23279.58 2384.64 16743.57 26735.95 4312.00
2449.75 64139.98 3656.96 0.00 56461.37 5125.00 17412.20 6382.82 6141.00 4784.70
4842.42 0.00 0.00 5631.00 675.00 0.00 4725.00 18193.63 5329.84 0.00

```

```

[59] 63179.18 3999.11 2423.29 55717.79 7679.00 0.00 648.00 4811.00 4247.00
541.00 34621.79 0.00 5619.00 651.00 4742.00 481.24 0.00 2457.66 17124.50
77371.46 0.00 29747.88 0.00 538.00 3451.00 6542.00 39681.10 5744.00 0.00

```

After Cleaning

OVERTIME_PAY	ALL_OTHER_PAY	TOTAL_PAY	CITY_RETIREMENT_CONTRIBUTIONS	BENEFIT_PAY	GENDER
23993.69	12596.92	128801.13	0.00	6760.77	MALE
41736.73	3918.57	111946.90	0.00	24845.07	MALE
36874.05	1525.00	171736.65	62468.67	19749.08	MALE
6242.67	2182.48	68982.84	0.00	18480.54	MALE
12196.36	880.26	62321.98	0.00	6443.85	MALE
31026.95	6545.00	200747.95	76447.96	19748.08	MALE
179.36	1494.22	100012.06	29167.19	19970.46	MALE

E. Analysis & Visualizations:

After cleaning the data, we did the below visualization based on different analysis questions.

Note: In order to use different libraries for good visualization, many packages were installed such as “gplot2”, “tidyverse”, “dplyr”, “ggforce”, etc.

1) What is the distribution of employees by gender in departments?

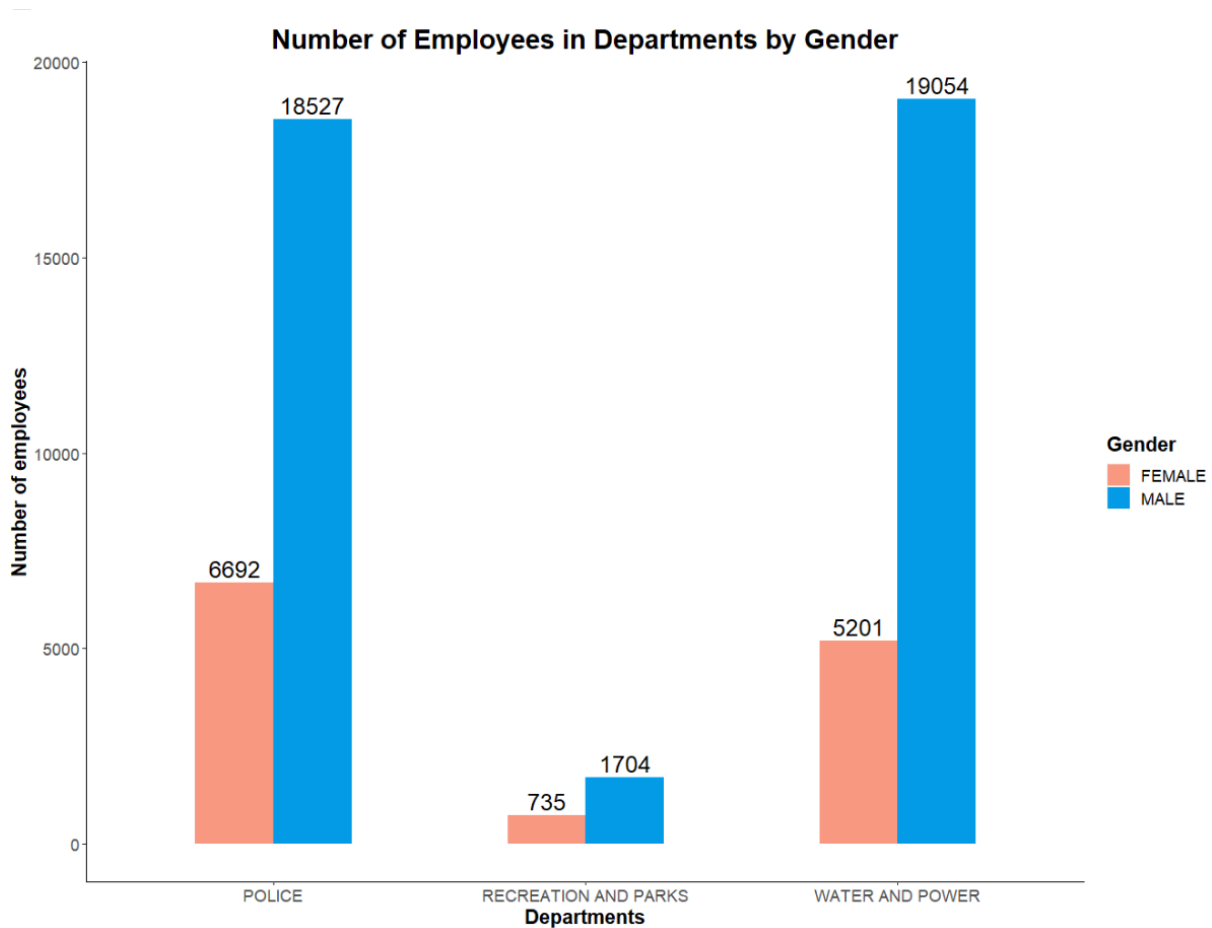


Figure1. Grouped bar chart showing the distribution of employees in departments

R features:

- **Plot Type:** Grouped bar plot(vertical)
- **Functions:** my_BarChartFun, aesthetics(aes), theme, geom_bar, labs, element_text, element_rect, element_line, scale_fill_manual, geom_text,
- **Libraries:** ggplot2

First, we installed the “ggplot2” package in the R studio

```
> install.packages("ggplot2")|
```

```
#load library ggplot2
library(ggplot2)

#store values in variables
Departments<-LA_Payroll_data$DEPARTMENT_TITLE
Gender<-LA_Payroll_data$GENDER

#write function of bar plot
my_BarChartFun<- function(mydataset,myxcol,myycol,mytitle,mycolor)
{
  ggplot(LA_Payroll_data, aes(x = Departments, fill = Gender)) +
    theme(axis.text.x=element_text(size=12),axis.text.y =element_text(size=12)) +
    geom_bar(position = position_dodge(),width = 0.5) +
    labs(title =mytitle, x=myxcol, y=myycol) +
    theme(axis.title.x = element_text(size = 15, face = "bold"), axis.title.y = element_text(size=15, face = "bold"),
          plot.title = element_text(size = 20, color = mycolor, hjust = 0.5, face="bold"), panel.background = element_rect("white"),
          axis.line = element_line(size = 0))+
    scale_fill_manual(values = c("#F89880","#039be5")) +
    geom_text(aes(label=..count..),stat = "count", position = position_dodge(0.5),vjust=-0.3, size=6) +
    theme(legend.title = element_text(size=15, face = "bold")) +
    theme(legend.text = element_text(size=12))
  |
}
my_BarChartFun(LA_Payroll_data, "Departments", "Number of employees", "Number of Employees in Departments by Gender","Black")
```

```

#Load library ggplot2
library(ggplot2)

#Store values in variables
Departments<-LA_Payroll_data$DEPARTMENT_TITLE
Gender<-LA_Payroll_data$GENDER

#Write function of bar plot
my_BarChartFun<- function(mydataset,myxcol,myycol,mytitle,mycolor)
{
  ggplot(LA_Payroll_data, aes(x = Departments, fill = Gender)) +
    theme(axis.text.x=element_text(size=12),axis.text.y =element_text(size=12)) +
    geom_bar(position = position_dodge(),width = 0.5) +
    labs(title =mytitle, x=myxcol, y=myycol) +
    theme(axis.title.x = element_text(size = 15, face = "bold"), axis.title.y =
    element_text(size=15, face = "bold"),
      plot.title = element_text(size = 20, color = mycolor, hjust = 0.5, face="bold"),
      panel.background = element_rect("white"),
      axis.line = element_line(size = 0))+
    scale_fill_manual(values = c("#F89880","#039be5")) +
    geom_text(aes(label=..count..),stat = "count", position = position_dodge(0.5),vjust=-
    0.3, size=6) +
    theme(legend.title = element_text(size=15, face = "bold")) +
    theme(legend.text = element_text(size=12))
}

my_BarChartFun(LA_Payroll_data, "Departments", "Number of employees", "Number of
Employees in Departments by Gender", "Black")

```

Insights:

Finding the department with the highest and lowest ratio of female-to-male employees was our first goal when we began to evaluate the data. According to the graph above, the ratio of men to women in the police department is 36%, compared to only 27% in the water and power department. Additionally, the ratio for the recreation and parks department is 43%, which is higher than that of other departments. It means despite having only 2439 employees, recreation and parks have the highest female-to-male employees' ratio, whereas water and power employs a much higher percentage of workers than other departments yet have the lowest female-to-male employee ratio. In addition, we can say from the above visual that majority of men are interested in the water and power profession, whereas majority of women are interested in working in the police department in Los Angeles.

2) What would be the distribution of pay in departments?

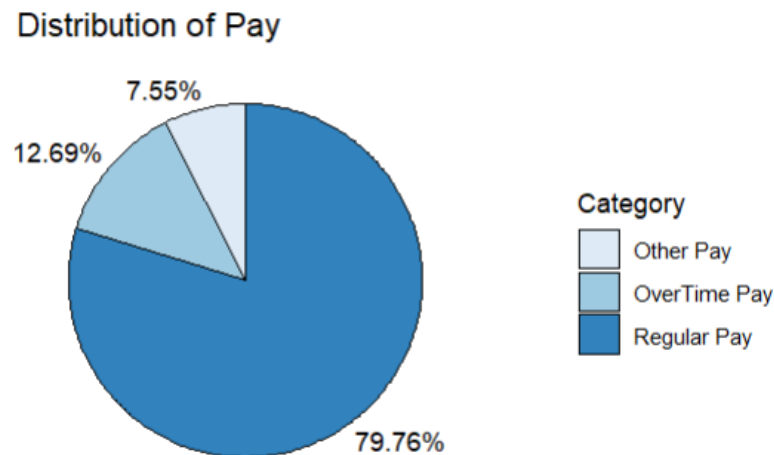


Figure2. Pie chart showing pay distribution of each department in percentage.

R features:

- **Plot Type:** Pie Chart
- **Functions:** geom_arc_bar, aesthetics(aes), geom_text, paste0, coord_fixed, scale_x_continuous, scale_y_continuous, scale_fill_brewer, theme_minimal, labs, theme_classic, theme, element_text, element_blank
- **Libraries:** ggplot2, dplyr, ggforce

In order to make a pie chart, the first thing we did was installed the three packages as shown below.

```
> install.packages("ggplot2")|
```

```
> install.packages("dplyr")|
```

```
> install.packages("ggforce")
```

```
# TotalPay
SUM_TOTAL_PAY = sum(LA_Payroll_data$TOTAL_PAY, na.rm = FALSE)

# OverTime
SUM_OVERTIME_PAY = sum(LA_Payroll_data$OVERTIME_PAY, na.rm = FALSE)

# OtherPay
Sum_ALL_OTHER_PAY = sum(LA_Payroll_data$ALL_OTHER_PAY, na.rm = FALSE)

# OtherTime %
OverTime_per = round((SUM_OVERTIME_PAY / SUM_TOTAL_PAY) * 100,2)

# OtherPay %
OtherPay_Per = round((Sum_ALL_OTHER_PAY / SUM_TOTAL_PAY) * 100,2)

# Calculating Regular Payment%
RegPay_per = 100 - OverTime_per - OtherPay_Per

# Create Dataframe
df <- data.frame(
  Pay=c("Regular Pay","OverTime Pay","Other Pay"),
  Percentage=c(RegPay_per,OverTime_per,OtherPay_Per)
)
df
```

```

# Install packages and using its library ggplot2,dplyr,ggforce
library(ggplot2)
library(dplyr)
library(ggforce)

# Set pie chart distribution
df <- df %>%
  mutate(end = 2 * pi * cumsum(Percentage)/sum(Percentage),
         start = lag(end, default = 0),
         middle = 0.5 * (start + end),
         hjust = ifelse(middle > pi, 1, 0),
         vjust = ifelse(middle < pi/2 | middle > 3 * pi/2, 0, 1))

# Pie chart with geom_arc_bar() to draw edges and arc
pie <- ggplot(df) +
  geom_arc_bar(aes(x0 = 0, y0 = 0, r0 = 0, r = 1,
                  start = start, end = end, fill = Pay))

pie <- pie +
  geom_text(aes(x = 1.05 * sin(middle),
               y = 1.05 * cos(middle),
               label = paste0(Percentage, "%"),
               hjust = hjust,
               vjust = vjust,

  )
  )

# Adjust so labels are not cut off for x Axis
pie <- pie +
  coord_fixed() +
  scale_x_continuous(limits = c(-1.5, 1.5),
                    name = "",
                    breaks = NULL,
                    labels = NULL

  )

```

```

# Adjust so labels are not cut off for Y Axis
pie <- pie +
  scale_y_continuous(limits = c(-1, 1.1),
                     name = "",
                     breaks = NULL,
                     labels = NULL)

# Add color scale (hex colors)
pie<- pie + scale_fill_brewer(palette="Blues")+
  theme_minimal()

# Remove labels and add title
pie <- pie +
  labs(x = NULL,
       y = NULL,
       fill = "Category",
       title = "Distribution of Pa")

# Adjust Title with Classic theme
pie <- pie +
  theme_classic() + theme(axis.line = element_blank(),
                          axis.text = element_blank(),
                          axis.ticks = element_blank(),
                          plot.title = element_text(hjust = 0.2, vjust =3.0)
  )
#Show pie chart
pie

```

```

# TotalPay
SUM_TOTAL_PAY = sum(LA_Payroll_data$TOTAL_PAY, na.rm = FALSE)

# OverTime
SUM_OVERTIME_PAY = sum(LA_Payroll_data$OVERTIME_PAY, na.rm = FALSE)

# OtherPay
Sum_ALL_OTHER_PAY = sum(LA_Payroll_data$ALL_OTHER_PAY, na.rm = FALSE)

# OtherTime %
OverTime_per = round((SUM_OVERTIME_PAY / SUM_TOTAL_PAY) * 100,2)

# OtherPay %
OtherPay_Per = round((Sum_ALL_OTHER_PAY / SUM_TOTAL_PAY) * 100,2)

# Caculating Regular Payment%
RegPay_per = 100 - OverTime_per -OtherPay_Per

```

```

# Create Dataframe
df <- data.frame(
  Pay=c("Regular Pay","OverTime Pay","Other Pay"),
  Percentage=c(RegPay_per,OverTime_per,OtherPay_Per)
)
df

# Install packages and using its library ggplot2,dplyr,ggforce
library(ggplot2)
library(dplyr)
library(ggforce)

# Set pie chart distribution
df <- df %>%
  mutate(end = 2 * pi * cumsum(Percentage)/sum(Percentage),
         start = lag(end, default = 0),
         middle = 0.5 * (start + end),
         hjust = ifelse(middle > pi, 1, 0),
         vjust = ifelse(middle < pi/2 | middle > 3 * pi/2, 0, 1))

# Pie chart with geom_arc_bar() to draw edges and arc
pie <- ggplot(df) +
  geom_arc_bar(aes(x0 = 0, y0 = 0, r0 = 0, r = 1,
                 start = start, end = end, fill = Pay))
pie <- pie +
  geom_text(aes(x = 1.05 * sin(middle),
               y = 1.05 * cos(middle),
               label = paste0(Percentage, "%"),
               hjust = hjust,
               vjust = vjust,
  )
)
)

```

```

# Adjust so labels are not cut off for X Axis
pie <- pie +
  coord_fixed() +
  scale_x_continuous(limits = c(-1.5, 1.5),
    name = "",
    breaks = NULL,
    labels = NULL
  )
# Adjust so labels are not cut off for Y Axis
pie <- pie +
  scale_y_continuous(limits = c(-1, 1.1),
    name = "",
    breaks = NULL,
    labels = NULL)
# Add color scale (hex colors)
pie<- pie + scale_fill_brewer(palette="Blues")+
  theme_minimal()
# Remove labels and add title
pie <- pie +
  labs(x = NULL,
    y = NULL,
    fill = "Category",
    title = "Distribution of Pa")
# Adjust Title with Classic theme
pie <- pie +
  theme_classic() + theme(axis.line = element_blank(),
    axis.text = element_blank(),
    axis.ticks = element_blank(),

```

```
        plot.title = element_text(hjust = 0.2, vjust = 3.0)
    )
#Show pie chart
pie
```

Insights:

When we first began to analyze our data set, Our main aim was to determine the distribution of pay(regular, other, and overtime) in these three departments, and from the above pie chart, we analyzed that these departments(Police, Recreation and Parks, Water and power) have a good distribution and payroll management as the regular pay is higher than the other and overtime pay which is 79.76%, and the overtime and other pay are 12.69% and 7.55% respectively. Moreover, according to California overtime law, ideally, overtime plus other pay should not be exceeded or equivalent to the regular pay and if they are higher or equal to regular pay, that means there is a huge fault in the distribution of manpower's hours.

3) Which ethnicity has the highest share of employment in all the departments?

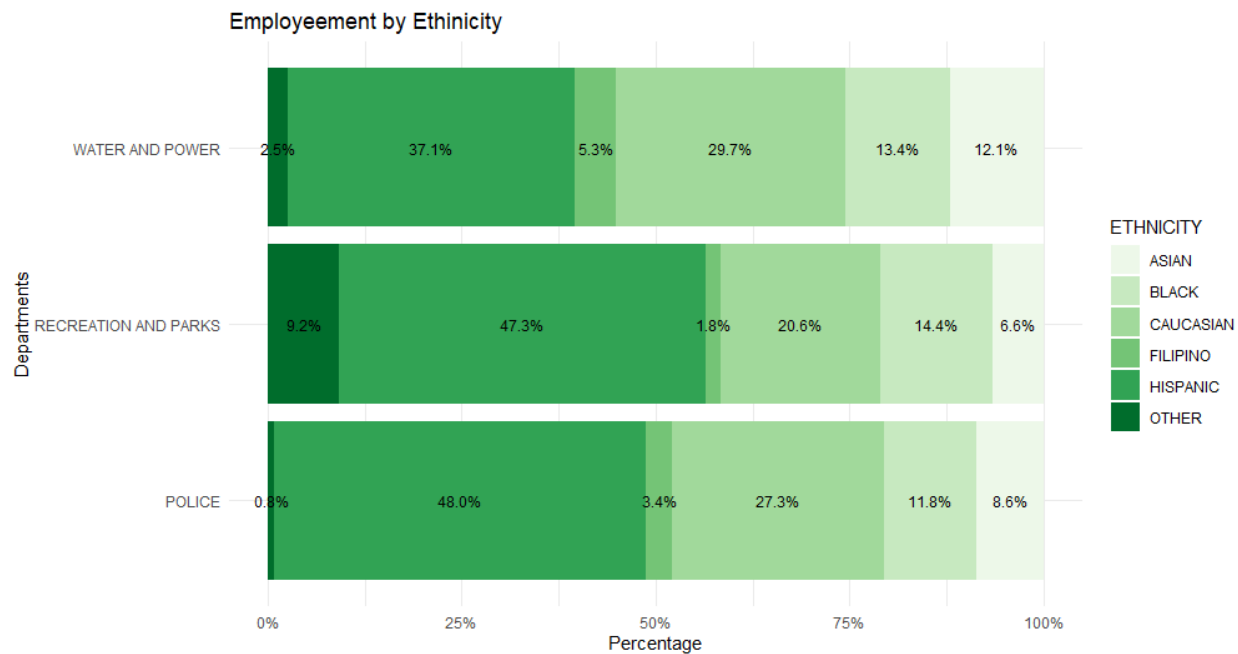


Figure3. Stacked 100% horizontal bar chart showing employment by ethnicity.

R features:

- **Plot Type:** Stacked 100% bar chart(Horizontal)
- **Functions:** group_by, summarise, aesthetics(aes), geom_bar, theme, element_text, geom_text, position_fill, scal_fill_brewer, theme_minimal, labs, scale_y_continuous, coord_flip
- **Libraries:** ggplot2, dplyr, ggthemes, scales, tidyverse, tidyr

First, we installed the below packages for the stacked 100% bar chart. Besides, “ggplot2” and “dplyr” also are required for this analysis but they are already installed earlier.

```
> install.packages("scales")|
```

```
> install.packages("ggthemes")
```

```
> install.packages("tidyverse")
```

```
> install.packages("tidyr")
```

```
library(ggplot2)
library(ggthemes)
library(scales)
library(dplyr)
library(tidyverse)
library(tidyr)

dt <- LA_Payroll_data %>%
  dplyr::group_by(DEPARTMENT_TITLE, ETHNICITY) %>%
  dplyr::tally() %>%
  dplyr::mutate(percent = (n / sum(n)))

dt

dt <- dt %>%
  mutate(ETHNICITY = ifelse(percent < 0.01, "OTHER", .$ETHNICITY))
dt

dt <- dt %>% group_by(DEPARTMENT_TITLE, ETHNICITY) %>%
  summarise(n = sum(n),
            percent = sum(percent),
            .groups = 'drop') %>%
  as.data.frame()
dt

p1 <-
  ggplot(data = dt, aes(x = DEPARTMENT_TITLE, y = n, fill = ETHNICITY))

p1 <- p1 + geom_bar(stat = "identity", position = "fill") +
  theme(axis.title = element_text(face = "bold"),
        legend.title = element_text(face = "bold"))

p1 <-
  p1 + geom_text(
    aes(label = paste0(sprintf(
      "%1.1f", percent * 100
    ), "%")),
    position = position_fill(vjust = 0.5),
    colour = "Black",
    size = 3
  ) +
  scale_fill_brewer(palette = "Greens") +
  theme_minimal()
p1
```



```

p1 <- p1 + theme_minimal(base_size = 10)
p1 <- p1 + labs(title = "Employement by Ethnicity")
p1 <- p1 + labs(x = "Departments", y = "Percentage")
p1 <- p1 + scale_y_continuous(labels = scales::percent)
p1
p1 <- p1 + coord_flip()
p1

```

```

library(ggthemes)
library(scales)
library(dplyr)
library(tidyverse)
library(tidyr)
dt <- LA_Payroll_data %>%
  dplyr::group_by(DEPARTMENT_TITLE, ETHNICITY) %>%
  dplyr::tally() %>%
  dplyr::mutate(percent = (n / sum(n)))
dt
dt <- dt %>%
  mutate(ETHNICITY = ifelse(percent < 0.01, "OTHER", .$ETHNICITY))
dt
dt <- dt %>% group_by(DEPARTMENT_TITLE, ETHNICITY) %>%
  summarise(n = sum(n),
            percent = sum(percent),
            .groups = 'drop') %>%
  as.data.frame()
dt

```

```

pl <-
  ggplot(data = dt, aes(x = DEPARTMENT_TITLE, y = n , fill = ETHNICITY))

pl <- pl + geom_bar(stat = "identity", position = "fill") +
  theme(axis.title = element_text(face = "bold"),
        legend.title = element_text(face = "bold"))
pl <-
  pl + geom_text(
    aes(label = paste0(sprintf(
      "%1.1f", percent * 100
    ), "%")),
    position = position_fill(vjust = 0.5),
    colour = "Black",
    size = 3
  ) +
  scale_fill_brewer(palette = "Greens") +
  theme_minimal()
pl
pl <- pl + theme_minimal(base_size = 10)
pl <- pl + labs(title = "Employeement by Ethnicity")
pl <- pl + labs(x = "Departments", y = "Percentage")
pl <- pl + scale_y_continuous(labels = scales::percent)
pl
pl <- pl + coord_flip()
pl

```

Insights:

Our initial approach was to analyze which ethnicity has the highest share of employment in all the departments and for better visualization, since we had more than 8 categories in the ethnicity field, we grouped the categories of ethnicity which has a value of less than 1% and counted it as others as you can see from the graph. After this grouping, we analyzed that “Hispanic” ethnicity has the highest share of employment in all the departments which is 37.1% for the water and power department, 47.3% for recreation and parks, and 48% for the police department. Apart from this we also analyzed that the police department has the highest Hispanic ethnicity among others which means 48% of workers had identified their ethnicity in this department.

F. Statistical Summary, Script, Function:

1) Statistical Summary:

The below screenshot shows the summary statistic of our complete dataset(LA_Payroll_data).

```
> summary(LA_Payroll_data)
  RECORD_NBR      PAY_YEAR DEPARTMENT_NO DEPARTMENT_TITLE  JOB_TITLE      EMPLOYMENT_TYPE
Length:51913   Min.   :2020   Min.   :70.00   Length:51913   Length:51913   Length:51913
Class :character 1st Qu.:2020   1st Qu.:70.00   Class :character Class :character Class :character
Mode  :character Median :2020   Median :88.00   Mode  :character Mode  :character Mode  :character
                Mean  :2020   Mean  :83.93
                3rd Qu.:2021   3rd Qu.:98.00
                Max.   :2021   Max.   :98.00

  JOB_STATUS      MOU_TITLE      REGULAR_PAY      OVERTIME_PAY      ALL_OTHER_PAY      TOTAL_PAY
Length:51913   Length:51913   Min.   : -6157   Min.   : -3527.2   Min.   : -63462   Min.   : 50008
Class :character Class :character 1st Qu.: 70370   1st Qu.:  953.4   1st Qu.:  2622   1st Qu.: 85947
Mode  :character Mode  :character Median : 93838   Median :  7180.8   Median :  6056   Median :117421
                Mean  : 96312   Mean  :15328.6   Mean  :  9117   Mean  :120757
                3rd Qu.:121143  3rd Qu.:21131.9  3rd Qu.:  9624   3rd Qu.:146808
                Max.   :398486   Max.   :265674.7   Max.   :275866   Max.   :479095

CITY_RETIREMENT_CONTRIBUTIONS BENEFIT_PAY      GENDER      ETHNICITY
Min.   : -313                Min.   :  0      Length:51913   Length:51913
1st Qu.:  622                1st Qu.:11607   Class :character Class :character
Median :  5734                Median :17866   Mode  :character Mode  :character
Mean   : 18102                Mean  :16015
3rd Qu.:31680                3rd Qu.:19549
Max.   :158436                Max.   :58786
```

1. Summary of Total-pay

The below screenshot depicts the summary statistic of total pay and here total pay is the addition of regular pay, overtime pay, and other pay.

```
> summary(LA_Payroll_data$TOTAL_PAY)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
50008  85947 117421 120757 146808  479095
> min(LA_Payroll_data$TOTAL_PAY)
[1] 50007.81
> max(LA_Payroll_data$TOTAL_PAY)
[1] 479095
> mean(LA_Payroll_data$TOTAL_PAY)
[1] 120756.8
> median(LA_Payroll_data$TOTAL_PAY)
[1] 117421.2
> sd(LA_Payroll_data$TOTAL_PAY)
[1] 45040.36
```

```

> summary(LA_Payroll_data$TOTAL_PAY)

  Min. 1st Qu.  Median    Mean 3rd Qu.   Max.
50008  85947 117421 120757 146808 479095

> min(LA_Payroll_data$TOTAL_PAY)

[1] 50007.81

> max(LA_Payroll_data$TOTAL_PAY)

[1] 479095

> mean(LA_Payroll_data$TOTAL_PAY)

[1] 120756.8

> median(LA_Payroll_data$TOTAL_PAY)

[1] 117421.2

> sd(LA_Payroll_data$TOTAL_PAY)

[1] 45040.36

```

Insights:

Based on the summary statistic of total pay, the least value of total pay was 50,008 and the most value was 4,79,095. Apart from this, the mean indicates the average value of total pay which is 1,20,756 and the median shows the middlemost value of the total pay when values are ordered from smallest to largest which is 1,17,421. As we can see from the summary, mean and median are close together which means the data set has a symmetrical distribution. Apart from this, the data point for 1st Qu. and 3rd Qu. from the median is quite similar so we can say that there is an equal dispersion among the smaller values and larger values of the dataset. Lastly, the standard

deviation is 45,040 which is high and indicates the value of the total pay field is spread out over a wide range.

2. Summary of Benefit pay:

The below screenshot shows the summary statistic of benefit payments.

```
> summary(LA_Payroll_data$BENEFIT_PAY)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0   11607   17866   16015   19549   58786
> min(LA_Payroll_data$BENEFIT_PAY)
[1] 0
> max(LA_Payroll_data$BENEFIT_PAY)
[1] 58786.03
> mean(LA_Payroll_data$BENEFIT_PAY)
[1] 16014.54
> median(LA_Payroll_data$BENEFIT_PAY)
[1] 17865.76
> sd(LA_Payroll_data$BENEFIT_PAY)
[1] 7242.321
```

```
> summary(LA_Payroll_data$BENEFIT_PAY)

  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0   11607   17866   16015   19549   58786

> min(LA_Payroll_data$BENEFIT_PAY)

[1] 0

> max(LA_Payroll_data$BENEFIT_PAY)

[1] 58786.03

> mean(LA_Payroll_data$BENEFIT_PAY)

[1] 16014.54

> median(LA_Payroll_data$BENEFIT_PAY)

[1] 17865.76

> sd(LA_Payroll_data$BENEFIT_PAY)

[1] 7242.321
```

Insights:

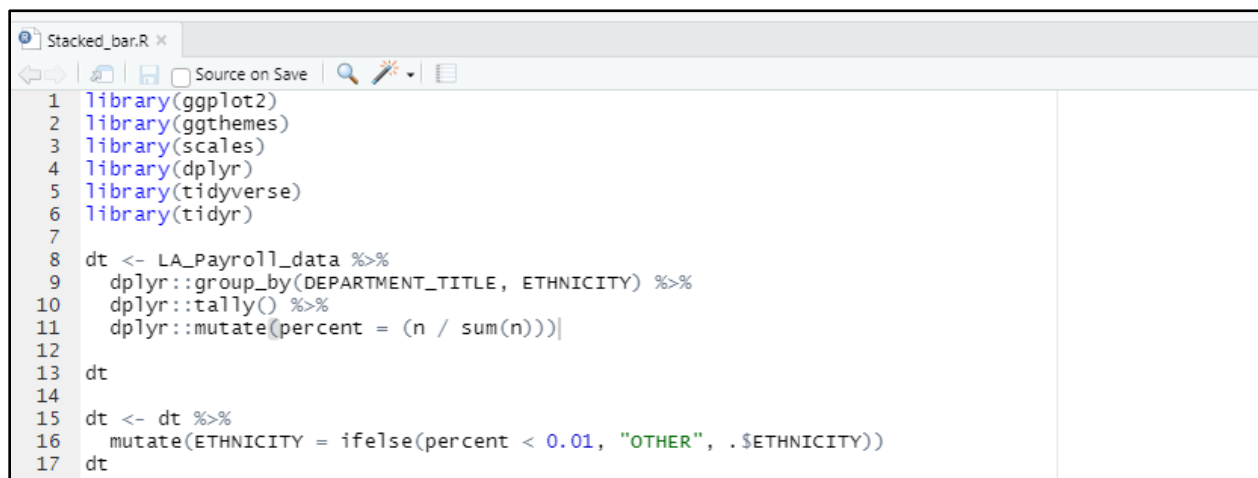
According to the summary statistic of benefit pay, the minimum value of benefit pay was 0 which indicates that many employees did not get any benefits from the departments. On the other hand, the maximum benefit pay value was 58,786. Moreover, the mean and median are 16,051 and 17,866 respectively which are quite close together which means our data set has a symmetrical distribution. Apart from this, we can see the data point for the 1st Qu. is further away from the median than the 3rd Qu. is from the median, so we can say that there is a greater dispersion among the smallest values of the dataset than among the larger values. And in the last, the standard deviation is 7,241 which is quite high which means the value of the benefit pay field is spread out over a wide range.

2) Script:

In order to create a script, we initially made the folder called “Script” saved the file as “Stacked_bar.R” . and set up the working directory in the R to run the script as shown below.

```
> setwd("~/Cal_State/5250/R Project/Scripts")
```

The below script is used to create a stacked bar chart.



```
1 library(ggplot2)
2 library(ggthemes)
3 library(scales)
4 library(dplyr)
5 library(tidyverse)
6 library(tidyr)
7
8 dt <- LA_Payroll_data %>%
9   dplyr::group_by(DEPARTMENT_TITLE, ETHNICITY) %>%
10  dplyr::tally() %>%
11  dplyr::mutate(percent = (n / sum(n)))
12
13 dt
14
15 dt <- dt %>%
16   mutate(ETHNICITY = ifelse(percent < 0.01, "OTHER", .$ETHNICITY))
17 dt
```

```

18
19 dt <- dt %>% group_by(DEPARTMENT_TITLE, ETHNICITY) %>%
20   summarise(n = sum(n),
21             percent = sum(percent),
22             .groups = 'drop') %>%
23   as.data.frame()
24 dt
25
26 p1 <-
27   ggplot(data = dt, aes(x = DEPARTMENT_TITLE, y = n, fill = ETHNICITY))
28
29
30 p1 <- p1 + geom_bar(stat = "identity", position = "fill") +
31   theme(axis.title = element_text(face = "bold"),
32         legend.title = element_text(face = "bold"))
33
34 p1 <-
35   p1 + geom_text(
36     aes(label = paste0(sprintf(
37       "%1.1f", percent * 100
38     ), "%")),
39     position = position_fill(vjust = 0.5),
40     colour = "Black",
41     size = 3
42   ) +
43   scale_fill_brewer(palette = "Greens") +
44   theme_minimal()
45 p1
46
47
48 p1 <- p1 + theme_minimal(base_size = 10)
49
50 p1 <- p1 + labs(title = "Employeement by Ethnicity")
51 p1 <- p1 + labs(x = "Departments", y = "Percentage")
52
53 p1 <- p1 + scale_y_continuous(labels = scales::percent)
54
55 p1
56 p1 <- p1 + coord_flip()
57 p1

```

```
library(ggthemes)
```

```
library(scales)
```

```
library(dplyr)
```

```
library(tidyverse)
```

```
library(tidyr)
```

```
dt <- LA_Payroll_data %>%
```

```
  dplyr::group_by(DEPARTMENT_TITLE, ETHNICITY) %>%
```



```

dplyr::tally() %>%

dplyr::mutate(percent = (n / sum(n)))

dt

dt <- dt %>%

  mutate(ETHNICITY = ifelse(percent < 0.01, "OTHER", .$ETHNICITY))

dt

dt <- dt %>% group_by(DEPARTMENT_TITLE, ETHNICITY) %>%

  summarise(n = sum(n),

            percent = sum(percent),

            .groups = 'drop') %>%

  as.data.frame()

dt

pl <-

  ggplot(data = dt, aes(x = DEPARTMENT_TITLE, y = n , fill = ETHNICITY))

pl <- pl + geom_bar(stat = "identity", position = "fill") +

  theme(axis.title = element_text(face = "bold"),

        legend.title = element_text(face = "bold"))

pl <-

pl + geom_text(

  aes(label = paste0(sprintf(

```

```

"% 1.1f", percent * 100

), "%"),

position = position_fill(vjust = 0.5),

colour = "Black",

size = 3

) +

scale_fill_brewer(palette = "Greens") +

theme_minimal()

pl

pl <- pl + theme_minimal(base_size = 10)

pl <- pl + labs(title = "Employeement by Ethnicity")

pl <- pl + labs(x = "Departments", y = "Percentage")

pl <- pl + scale_y_continuous(labels = scales::percent)

pl

pl <- pl + coord_flip()

pl

```

3) Function:

The below function(my_BarChartFun) is used to create grouped bar plot for the analysis – **distribution of employees by gender in departments**. We have passed the input of the data frame name, the x column, the y column, the title of the plot, and the color in the parameter. When we called the function, we passed the arguments that we wanted to show in the grouped bar graph. And in order to run this function we did set up the working directory in R as below.

```
> setwd("~/Cal_State/5250/R Project/Scripts")
```

```
#write function of bar plot
my_BarChartFun<- function(mydataset,myxcol,myycol,mytitle,mycolor)
{
  ggplot(LA_Payroll_data, aes(x = Departments, fill = Gender)) +
    theme(axis.text.x=element_text(size=12),axis.text.y =element_text(size=12)) +
    geom_bar(position = position_dodge(),width = 0.5) +
    labs(title =mytitle, x=myxcol, y=myycol) +
    theme(axis.title.x = element_text(size = 15, face = "bold"), axis.title.y = element_text(size=15, face = "bold"),
          plot.title = element_text(size = 20, color = mycolor, hjust = 0.5, face="bold"), panel.background = element_rect("white"),
          axis.line = element_line(size = 0))+
    scale_fill_manual(values = c("#F89880","#039be5")) +
    geom_text(aes(label=..count..),stat = "count", position = position_dodge(0.5),vjust=-0.3, size=6) +
    theme(legend.title = element_text(size=15, face = "bold")) +
    theme(legend.text = element_text(size=12))
}
my_BarChartFun(LA_Payroll_data, "Departments", "Number of employees", "Number of Employees in Departments by Gender","Black")
```

```
#write function of bar plot

my_BarChartFun<- function(mydataset,myxcol,myycol,mytitle,mycolor) {

  ggplot(LA_Payroll_data, aes(x = Departments, fill = Gender)) +

    theme(axis.text.x=element_text(size=12),axis.text.y =element_text(size=12)) +

    geom_bar(position = position_dodge(),width = 0.5) +

    labs(title =mytitle, x=myxcol, y=myycol) +

    theme(axis.title.x = element_text(size = 15, face = "bold"), axis.title.y =
    element_text(size=15, face = "bold"),
```

```

plot.title = element_text(size = 20, color = mycolor, hjust = 0.5, face="bold"),
panel.background = element_rect("white"),

axis.line = element_line(size = 0))+

scale_fill_manual(values = c("#F89880", "#039be5")) +

geom_text(aes(label=..count..), stat = "count", position = position_dodge(0.5), vjust=-
0.3, size=6) +

theme(legend.title = element_text(size=15, face = "bold")) +

theme(legend.text = element_text(size=12))

}

my_BarChartFun(LA_Payroll_data, "Departments", "Number of employees", "Number of
Employees in Departments by Gender", "Black")

```

References:

- Green, J., & Stecker, T. (2022, August 30). *California passes law requiring companies like Meta, Disney to post salary range*. Bloomberg.com. Retrieved December 15, 2022, from <https://www.bloomberg.com/news/articles/2022-08-30/california-passes-law-requiring-companies-like-meta-disney-to-post-salary-range?leadSource=uverify+wall/>
- Liberto, D. (2022, September 13). *What is a quartile? example and how it works*. Investopedia. Retrieved December 15, 2022, from <https://www.investopedia.com/terms/q/quartile.asp/>