Assignment 4 - Anshita Thakkar
Visual Analytics – Professor Emily Ko
Data Manipulation, Filtering and Visualization

Create **FIVE Different** subsets of SIVA data or processed SIVA data using the data manipulating or filtering R functions (e.g., filter, select, ifelse, group_by, summarise) that were discussed at the class meeting. Using head function, show the first 10 observations of each of the five data sets in addition to the R syntax used to create the data.

First, I read in the csv file:

```
siva = read.csv("/Users/anshitathakkar/Documents/Visual analytics/Week 6 - Data manipulation, filtering and visualiations/Assignment 4/siva.csv")
siva
```

Then, I examined the structure to see the numeric and character variables to aid my analyses and graph plotting decisions.

```
str(siva)
```

```
'data.frame':   53815 obs. of  29 variables:
 $ xgra_n1clb_nbr           : int  51407 23460 53417 14382 40539 53945 35983 43669 29279 14254 ...
 $ Siva_Rental_Number       : int  67041 56084 70279 15105 49797 71102 43104 54673 33940 14950 ...
 $ rent_area_loc            : int  156 204 181 1515 259 165 177 276 2167 953 ...
 $ Date_of_Survey           : chr  "5/18/2011" "2/5/2011" "6/14/2011" "1/4/2010" ...
 $ Day_of_Week              : chr  "Wednesday" "Saturday" "Tuesday" "Monday" ...
 $ Time                     : chr  "7:48:30" "22:06:37" "5:35:48" "23:58:56" ...
 $ Survey_Type              : chr  "SV Web Sol." "SV Web Sol." "SV Web Sol." "SV Web Sol." ...
 $ Purpose_of_Rental        : chr  "Bus." "Leis. / Pers." "Bus." "Leis. / Pers." ...
 $ Recom_mend_Siva          : int  8 8 8 7 9 9 9 6 9 5 ...
 $ Staff_Courtesy           : int  9 8 7 8 9 9 9 9 8 7 ...
 $ Speed_of_Service         : int  8 8 8 7 9 9 9 8 9 5 ...
 $ Veh_Equip_Condition      : int  9 5 8 8 9 9 9 9 8 8 ...
 $ Trans_Billing_as_Expected: int  9 8 8 8 9 9 9 9 9 7 ...
 $ Value_for_the_Money      : int  9 7 8 8 9 9 9 6 7 7 ...
 $ Area                     : chr  "01602 - LOVE FIELD AP TX" "07286 - VALLEJO CA OAP" "01850 - RICHMOND VA AP" "05743 - PICO CA OAP" ...
 $ loc_nm                   : chr  "DALLAS LOVE FIELD" "VALLEJO HLE" "RICHMOND INTL AP" "PICO HLE" ...
 $ ga_region_desc           : chr  "SOUTHWEST REGION" "WESTERN REGION" "MID ATLANTIC REGION" "WESTERN REGION" ...
 $ xgra_ckot_ts             : chr  "5/15/2011" "1/31/2011" "6/12/2011" "12/23/2009" ...
 $ xgra_ckin_ts             : chr  "5/17/2011" "2/3/2011" "6/13/2011" "1/3/2010" ...
 $ xgra_vclass_reserv       : chr  "C" "A" "F" "D" ...
 $ xgra_veh_class           : chr  "Q4" "B" "YF" "YF" ...
 $ rent_loc_type            : chr  "AP" "OFF AP" "AP" "OFF AP" ...
 $ cust_tier_code           : chr  "FG" "N1" "RG" "RG" ...
 $ booking_channel_code     : chr  "SIVA.COM" "SIVA.COM" "SIVA.COM" "SIVA.COM" ...
 $ col34_total_charges      : num  247.3 128 75.8 468.5 42.8 ...
 $ col38_currency           : chr  "USD" "USD" "USD" "USD" ...
 $ Total_charge_USD         : num  247.3 128 75.8 468.5 42.8 ...
 $ Survey_checkout_diff     : int  2 3 2 2 2 4 2 2 6 2 ...
 $ booking_channel_dummy    : int  1 1 1 1 1 0 0 0 1 1 ...
```

#1 Create a Variable Using Nested Conditional Statement

```
siva$total <- (siva$col34_total_charges + siva$Total_charge_USD)
summary(siva$total)
head(siva,10)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    0.0   173.8   317.2   450.9   534.0 28557.2 
```

col34_total_charges <dbl>	col38_currency <chr>	Total_charge_USD <dbl>	Survey_checkout_diff <int>	booking_channel_dummy <int>	total <dbl>
247.29	USD	247.29	2	1	494.58
128.04	USD	128.04	3	1	256.08
75.85	USD	75.85	2	1	151.70
468.51	USD	468.51	2	1	937.02
42.84	USD	42.84	2	1	85.68
107.92	USD	107.92	4	0	215.84
224.68	USD	224.68	2	0	449.36
387.63	USD	387.63	2	0	775.26
45.08	USD	45.08	6	1	90.16
229.10	USD	229.10	2	1	458.20

#2 ifelse function - If less than or equal 173.8, low, if between 317.2 and 534.0, avergae,otherwise high

```
siva$charges <- ifelse(siva$total <= 173.8, "low", ifelse(siva$total <= 534.0, "average", "high"))

siva %>% head(10)
```

col34_total_charges <dbl>	col38_currency <chr>	Total_charge_USD <dbl>	Survey_checkout_diff <int>	booking_channel_dummy <int>	total <dbl>	charges <chr>
247.29	USD	247.29	2	1	494.58	average
128.04	USD	128.04	3	1	256.08	average
75.85	USD	75.85	2	1	151.70	low
468.51	USD	468.51	2	1	937.02	high
42.84	USD	42.84	2	1	85.68	low
107.92	USD	107.92	4	0	215.84	average
224.68	USD	224.68	2	0	449.36	average
387.63	USD	387.63	2	0	775.26	high
45.08	USD	45.08	6	1	90.16	low
229.10	USD	229.10	2	1	458.20	average

```
siva$speed_category <- ifelse(siva$Speed_of_Service > 5, "Fast", "Slow")

head(siva,10)
```

Description: df [10 × 32]

col34_total_charges <dbl>	col38_currency <chr>	Total_charge_USD <dbl>	Survey_checkout_diff <int>	booking_channel_dummy <int>	total <dbl>	charges <chr>	speed_category <chr>
247.29	USD	247.29	2	1	494.58	average	Fast
128.04	USD	128.04	3	1	256.08	average	Fast
75.85	USD	75.85	2	1	151.70	low	Fast
468.51	USD	468.51	2	1	937.02	high	Fast
42.84	USD	42.84	2	1	85.68	low	Fast
107.92	USD	107.92	4	0	215.84	average	Fast
224.68	USD	224.68	2	0	449.36	average	Fast
387.63	USD	387.63	2	0	775.26	high	Fast
45.08	USD	45.08	6	1	90.16	low	Fast
229.10	USD	229.10	2	1	458.20	average	Slow

#3 Filter

```
library(dplyr)

siva %>% filter(Staff_Courtesy >= 7) %>% head(10)
```

Description: df [10 × 31]

Staff_Courtesy <int>	Speed_of_Service <int>	Veh_Equip_Condition <int>	Trans_Billing_as_Expected <int>	Value_for_the_Money <int>	Area <chr>
9	8	9	9	9	01602 – LOVE FIELD AP TX
8	8	5	8	7	07286 – VALLEJO CA OAP
7	8	8	8	8	01850 – RICHMOND VA AP
8	7	8	8	8	05743 – PICO CA OAP
9	9	9	9	9	07787 – NEWPORT RI OAP
9	9	9	9	9	01450 – ATLANTA AP GA
9	9	9	9	9	05426 – ELIZABETHTOWN: KY OAP
9	8	9	9	6	02170 – SALT LAKE CITY UT AP
8	9	8	9	7	07275 – WALSH RD. SANTA CLARA CA OAP
7	5	8	7	7	07781 – BEDFORD MA OAP

```
siva %>% filter(col38_currency == "CAD") %>% head(10)
```

booking_channel_code <chr>	col34_total_charges <dbl>	col38_currency <chr>	Total_charge_USD <dbl>	Survey_checkout_diff <int>	booking_channel_dummy <int>	total <dbl>	charges <chr>
WALKUP	234.86	CAD	223.42	5	0	458.28	average
800#	171.64	CAD	163.07	2	0	334.71	average
SIVA.COM	194.63	CAD	184.23	6	1	378.86	average
SIVA.COM	199.15	CAD	189.21	2	1	388.36	average
SIVA.COM	136.72	CAD	129.30	2	1	266.02	average
SIVA.COM	110.91	CAD	105.49	2	1	216.40	average
GDSB	1027.62	CAD	975.76	5	0	2003.38	high
SIVA.COM	100.32	CAD	95.02	3	1	195.34	average
800#	59.93	CAD	56.78	2	0	116.71	low
SIVA.COM	136.76	CAD	129.58	2	1	266.34	average

1–10 of 10 rows | 25–32 of 31 columns

#4 Group by and Summarize

```
siva %>%
  group_by(booking_channel_code) %>%
  summarise(Total_charge_USD = mean(Total_charge_USD)) %>%
  head(10)
```

A tibble: 10 × 2

booking_channel_code <chr>	Total_charge_USD <dbl>
800#	252.0832
GDSA	199.0619
GDSB	202.3806
GDSC	204.0189
GDSS	268.7908
LINK FR INS CO	326.9616
LOCAL RES	229.4357
O	195.3187
OTHER INTERNET	270.5851
SIVA.COM	230.2199

1–10 of 10 rows

```
siva %>%
  group_by(loc_nm) %>%
  summarise(Total_charge_USD = mean(Total_charge_USD)) %>%
  head(10)
```

A tibble: 10 × 2

loc_nm <chr>	Total_charge_USD <dbl>
10TH STREET HLE	249.95190
12 EAST 13TH ST (E)	237.04943
126 W. 55TH ST. (W)	437.64025
17TH & BEN FRANKLIN HLE	81.80885
210 W. 77TH ST. (W)	212.03111
214 W. 95TH ST. (W)	236.41289
222 E. 40TH ST. (E)	276.58500
270 W 60TH ST	182.57250
310 E. 48TH ST. (E)	190.33625
323 W. 34TH ST. (W)	190.33750

1–10 of 10 rows

```
siva %>%
  group_by(ga_region_desc) %>%
  summarise(Total_charge_USD = mean(Total_charge_USD)) %>%
  head(10)
```

ga_region_desc <chr>	Total_charge_USD <dbl>
CA LICENSEES	200.5451
CANADA REGION	240.3415
CENTRAL REGION	197.8417
MID ATLANTIC REGION	196.1271
NORTHEAST REGION	237.3571
PUERTO RICO VIRGIN ISLANDS	291.8128
SOUTHEAST REGION	233.2493
SOUTHWEST REGION	214.8488
US LICENSEES	218.8760
WEST CENTRAL REGION	263.7556

1–10 of 10 rows

#5 Select

```
siva %>% select(booking_channel_code, col38_currency) %>% head(10)
```

Description: df [10 × 2]

	booking_channel_code <chr>	col38_currency <chr>
1	SIVA.COM	USD
2	SIVA.COM	USD
3	SIVA.COM	USD
4	SIVA.COM	USD
5	SIVA.COM	USD
6	800#	USD
7	LOCAL RES	USD
8	GDSB	USD
9	SIVA.COM	USD
10	SIVA.COM	USD

1–10 of 10 rows

```
siva %>% select(charges, Purpose_of_Rental) %>% head(10)
```

Description: df [10 × 2]

	charges <chr>	Purpose_of_Rental <chr>
1	average	Bus.
2	average	Leis. / Pers.
3	low	Bus.
4	high	Leis. / Pers.
5	low	Leis. / Pers.
6	average	Bus.
7	average	Bus.
8	high	Bus.
9	low	Bus.
10	average	Bus.

1–10 of 10 rows

#6 Arrange

```
siva%>% arrange(Purpose_of_Rental)
```

Description: df [53,815 × 32]

xgra_n1clb_nbr <int>	Siva_Rental_Number <int>	rent_area_loc <int>	Date_of_Survey <chr>	Day_of_Week <chr>	Time <chr>	Survey_Type <chr>	Purpose_of_Rental <chr>	Recom_mend_Siva <int>
51407	67041	156	5/18/2011	Wednesday	7:48:30	SV Web Sol.	Bus.	8
53417	70279	181	6/14/2011	Tuesday	5:35:48	SV Web Sol.	Bus.	8
53945	71102	165	6/29/2011	Wednesday	5:34:06	SV Web Sol.	Bus.	9
35983	43104	177	9/23/2010	Thursday	6:24:40	SV Web Sol.	Bus.	9
43669	54673	276	1/23/2011	Sunday	9:48:08	SV Web Sol.	Bus.	6
29279	33940	2167	7/6/2010	Tuesday	14:38:54	SV Web Sol.	Bus.	9
14254	14950	953	1/7/2010	Thursday	5:49:29	SV Web Sol.	Bus.	5
31288	36540	160	8/2/2010	Monday	10:37:46	SV Web Sol.	Bus.	9
30328	35319	371	7/21/2010	Wednesday	12:31:16	SV Web Sol.	Bus.	7
23011	25652	242	4/13/2010	Tuesday	7:40:10	SV Web Sol.	Bus.	4

1–10 of 53,815 rows | 1–9 of 32 columns

Previous 1 2 3 4 5 6 … 100 Next

```
siva%>% arrange(charges, speed_category)

head(siva, 10)
```

Description: df [53,815 × 32]

col34_total_charges <dbl>	col38_currency <chr>	Total_charge_USD <dbl>	Survey_checkout_diff <int>	booking_channel_dummy <int>	total <dbl>	charges <chr>	speed_category <chr>
247.29	USD	247.29	2	1	494.58	average	Fast
128.04	USD	128.04	3	1	256.08	average	Fast
107.92	USD	107.92	4	0	215.84	average	Fast
224.68	USD	224.68	2	0	449.36	average	Fast
259.35	USD	259.35	5	0	518.70	average	Fast
100.89	USD	100.89	2	0	201.78	average	Fast
133.76	USD	133.76	2	1	267.52	average	Fast
173.91	USD	173.91	14	0	347.82	average	Fast
204.51	USD	204.51	2	0	409.02	average	Fast
179.45	USD	179.45	2	0	358.90	average	Fast

1–10 of 53,815 rows | 25–32 of 32 columns

Previous 1 2 3 4 5 6 … 100 Next

#7 Mutate

```
siva%>%
 mutate(TOTAL = col34_total_charges + Total_charge_USD)%>%
 head(10)
```

Description: df [10 × 33]

col34_total_charges <dbl>	col38_currency <chr>	Total_charge_USD <dbl>	Survey_checkout_diff <int>	booking_channel_dummy <int>	total <dbl>	charges <chr>	speed_category <chr>	TOTAL <dbl>
247.29	USD	247.29	2	1	494.58	average	Fast	494.58
128.04	USD	128.04	3	1	256.08	average	Fast	256.08
75.85	USD	75.85	2	1	151.70	low	Fast	151.70
468.51	USD	468.51	2	1	937.02	high	Fast	937.02
42.84	USD	42.84	2	1	85.68	low	Fast	85.68
107.92	USD	107.92	4	0	215.84	average	Fast	215.84
224.68	USD	224.68	2	0	449.36	average	Fast	449.36
387.63	USD	387.63	2	0	775.26	high	Fast	775.26
45.08	USD	45.08	6	1	90.16	low	Fast	90.16
229.10	USD	229.10	2	1	458.20	average	Slow	458.20

1–10 of 10 rows | 26–34 of 33 columns

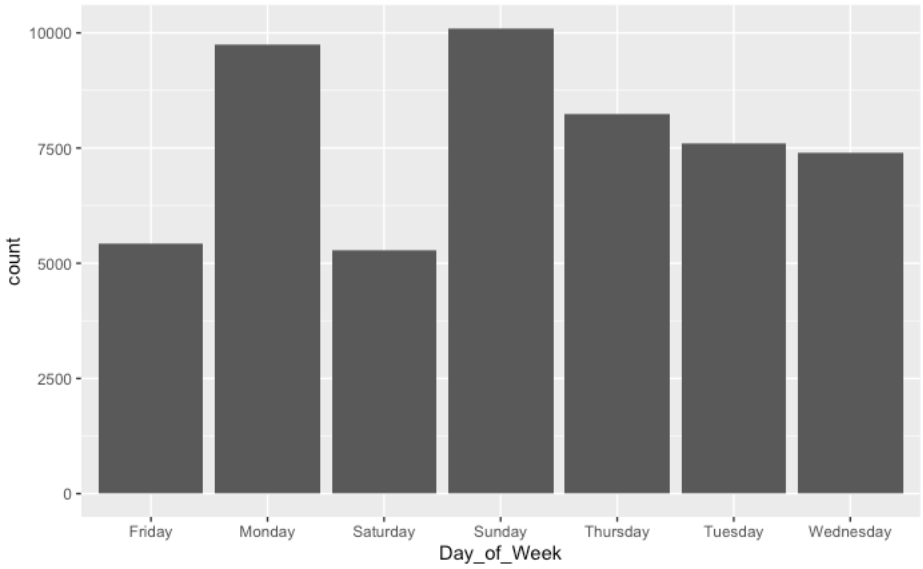
With SIVA data, create **FIVE Different** visualizations that were discussed at the class meeting. Using **Export function under Plots in R Studio**, export the plot and present it in the pdf file. Tell at least ONE interesting story (i.e., finding) from each of the visualizations.

Example)

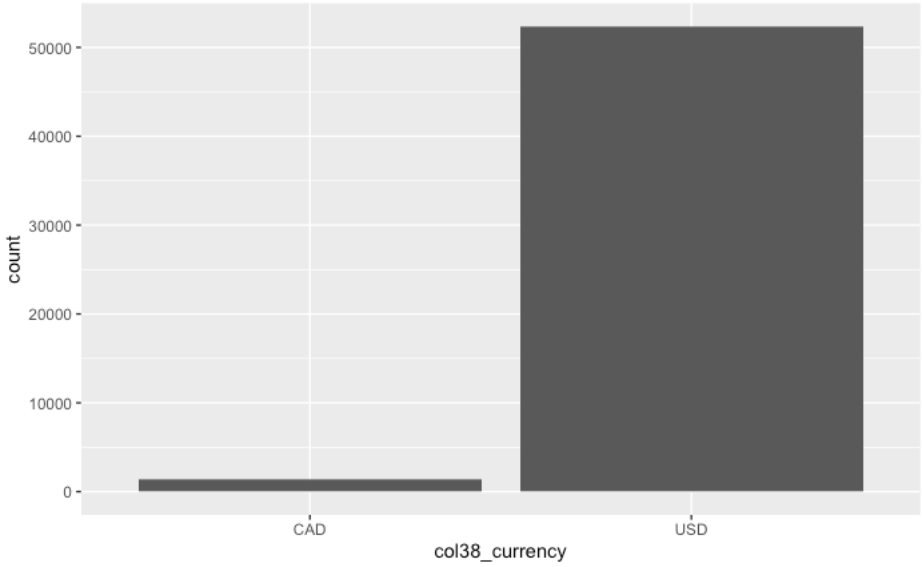
Code)

#1 Basic barplot

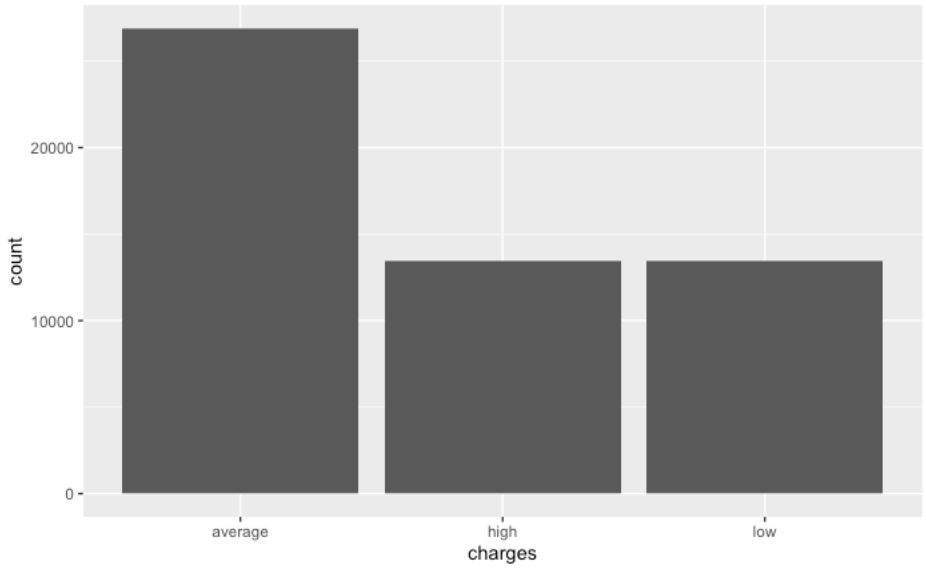
```
ggplot(data=siva, aes(x=Day_of_Week)) + geom_bar(stat="count")
```




```
ggplot(data=siva, aes(x=col38_currency)) + geom_bar(stat="count")
```

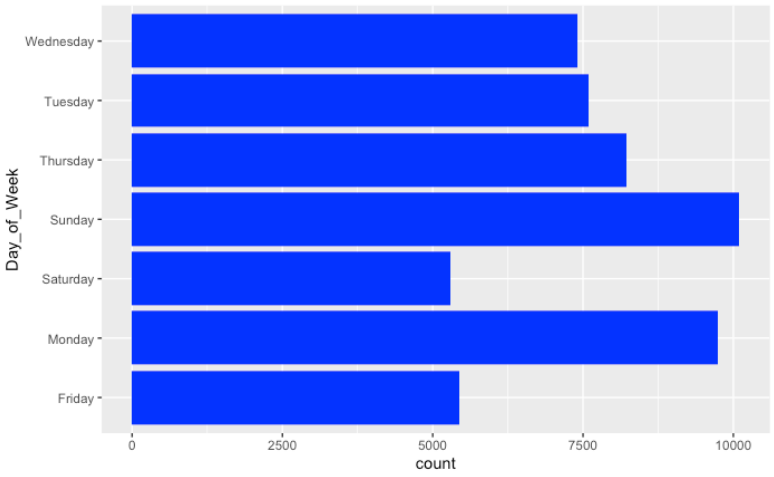



```
ggplot(data=siva, aes(x=charges)) + geom_bar(stat="count")
```

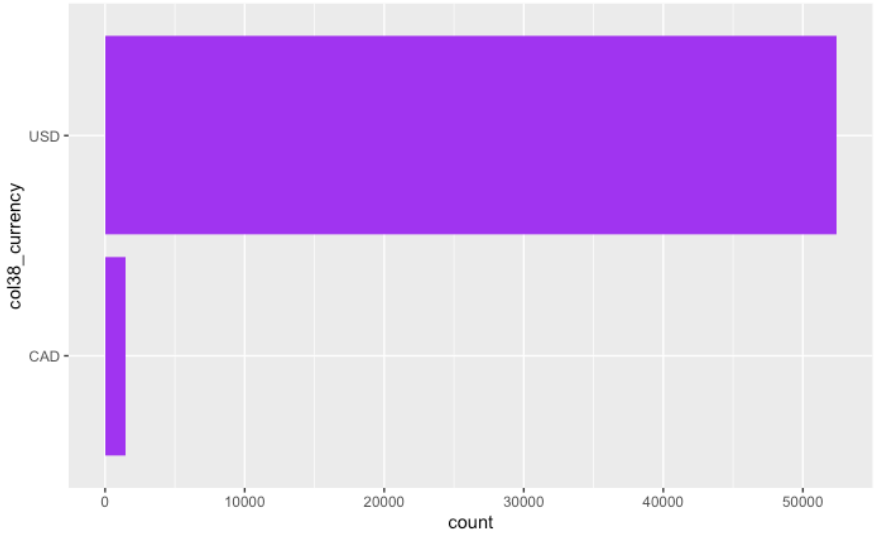


#change colours & coord flip

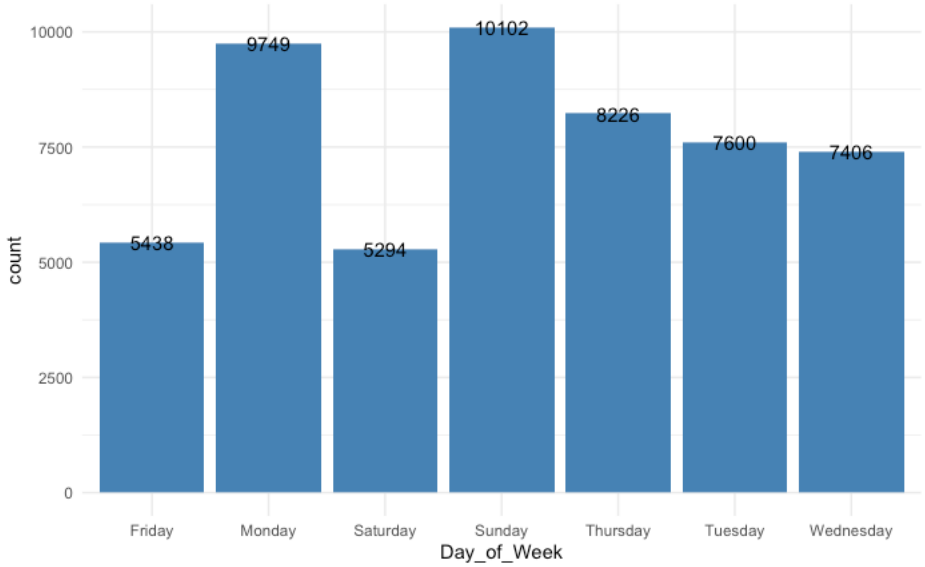
```
ggplot(data=siva, aes(x=Day_of_Week)) + geom_bar(fill="blue") + coord_flip()
```




```
ggplot(data=siva, aes(x=col38_currency)) + geom_bar(fill="purple") + coord_flip()
```



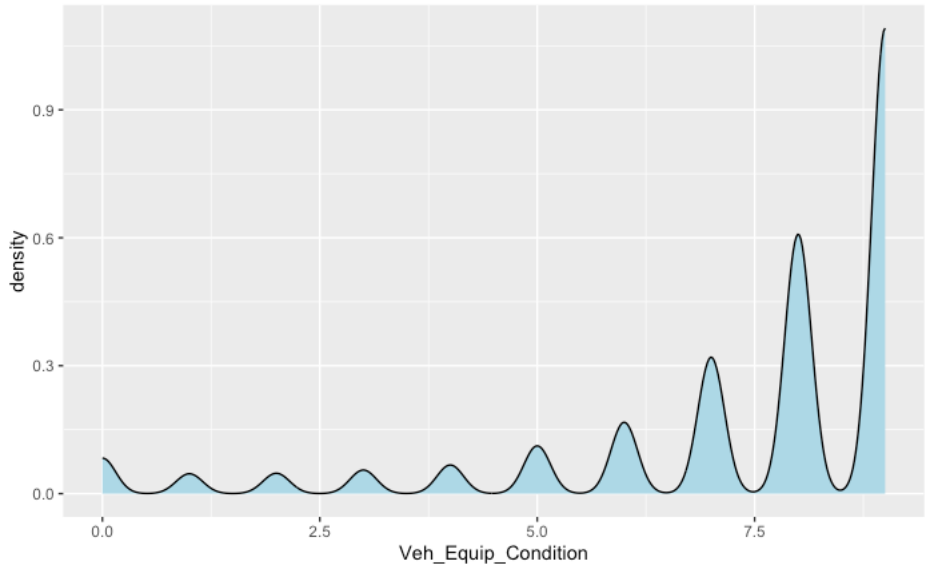

```
ggplot(data=siva, aes(x=Day_of_Week)) +
geom_bar(fill="steelblue") +
geom_text(stat = "count", aes(label = after_stat(count)),
color="black") + theme_minimal()
```



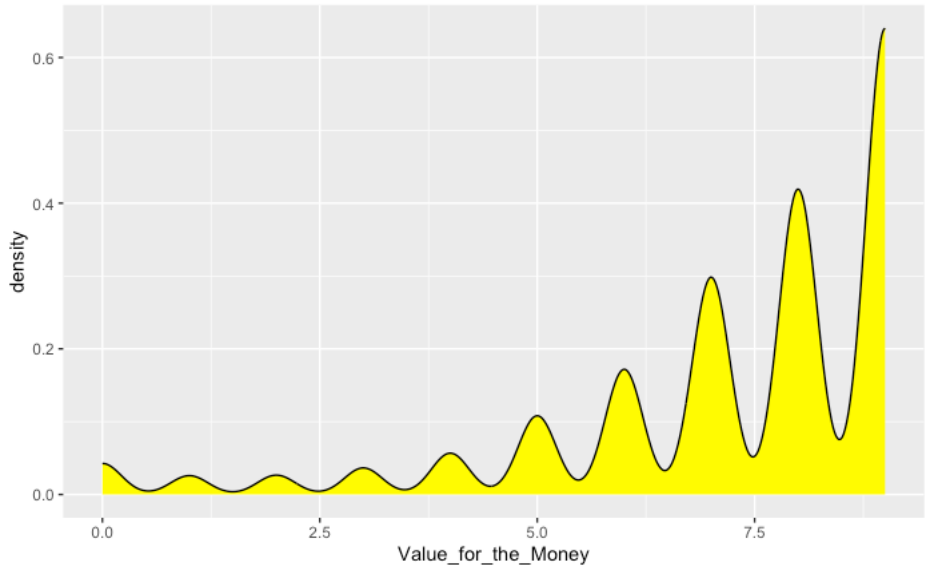
A bar plot is a graphical representation of categorical data that displays the frequency of a variable and provides a visual summary of the distribution data in a dataset. This makes it easy to compare and understand variables in a dataset. An interesting finding is that most customers paid in US dollars as opposed to Canadian dollars. Also, the day of the weeks that most surveys were completed were Sunday.

#2 density plot

```
ggplot(siva, aes(x = Veh_Equip_Condition)) +
  geom_density(fill = "lightblue")
```



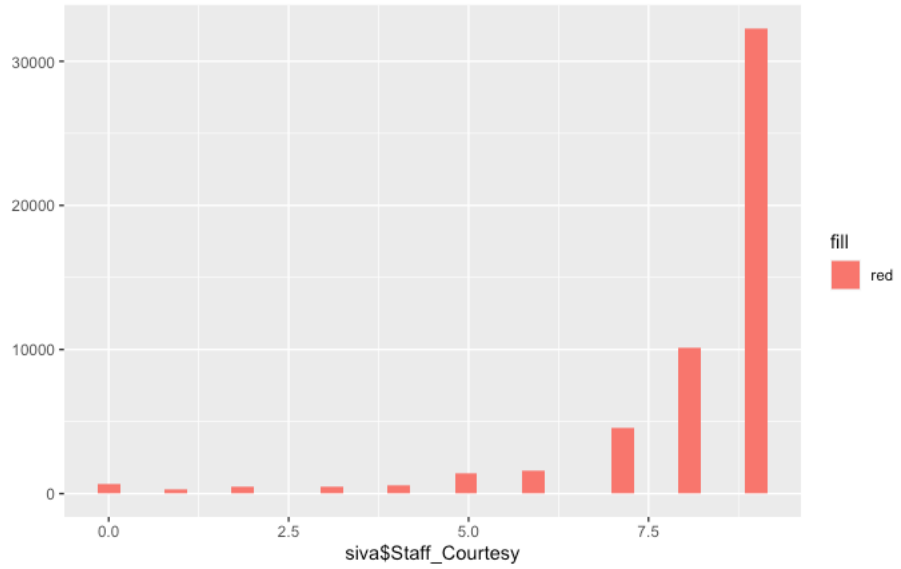

```
ggplot(siva, aes(x = Value_for_the_Money)) +
  geom_density(fill = "yellow")
```



A density plot observes the distribution of a variable in a dataset. The Vehicle Equipment Condition and Value for Money plot both show that overall the equipment quality and value for money were rated highly, and overall customers were largely satisfied. Further analysis would lead to more detailed extrapolation of the results.

#3 Histogram

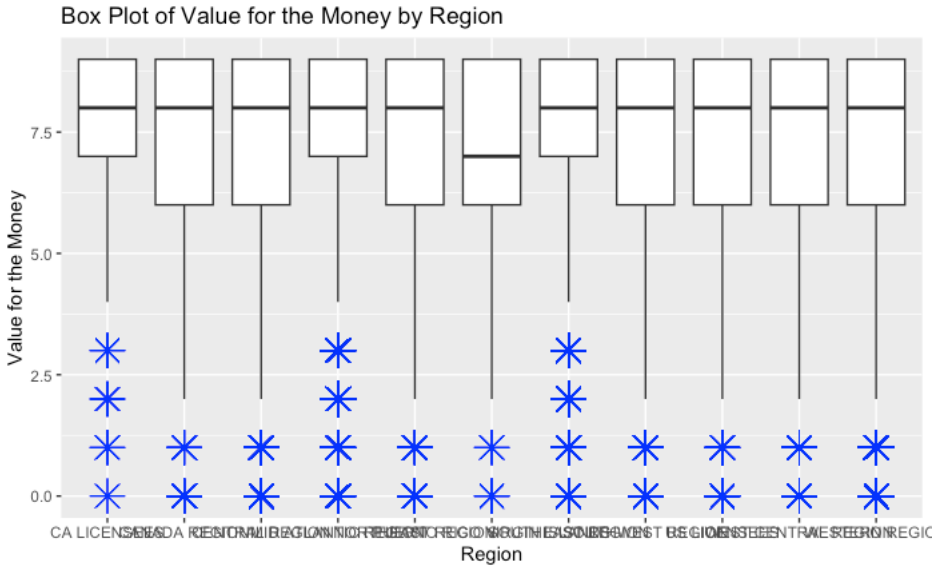
qplot(siva$Staff_Courtesy, fill = "red")



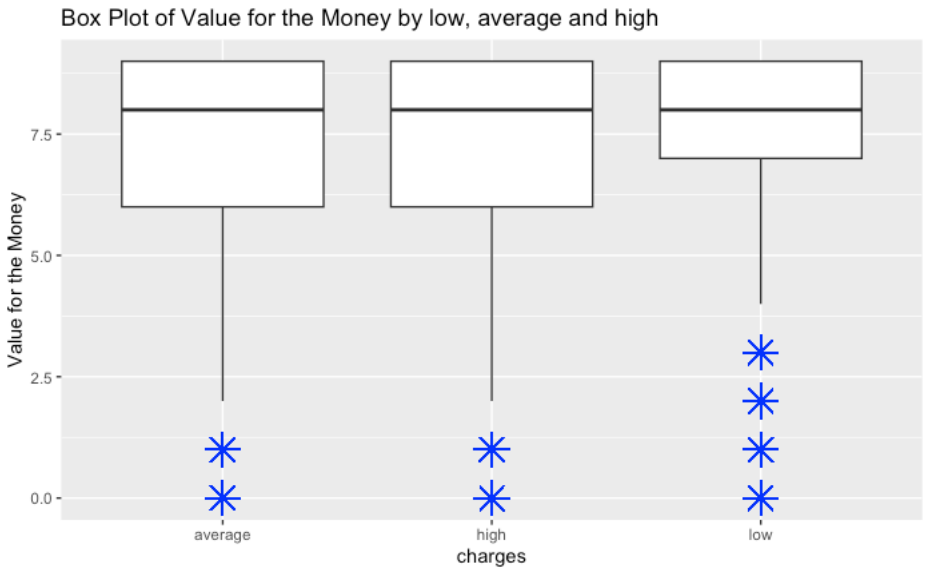
A histogram is a graphical representation of a continuous or discrete dataset. Data is divided into bins. Overall, the staff courtesy score was high indicating that staff were courteous rather than non-courteous.

#4 box plot

```
ggplot(siva, aes(x = ga_region_desc, y = Value_for_the_Money)) +
  geom_boxplot(outlier.color = "blue", outlier.shape = 8, outlier.size = 6) +
  labs(title = "Box Plot of Value for the Money by Region",
       x = "Region",
       y = "Value for the Money")
```

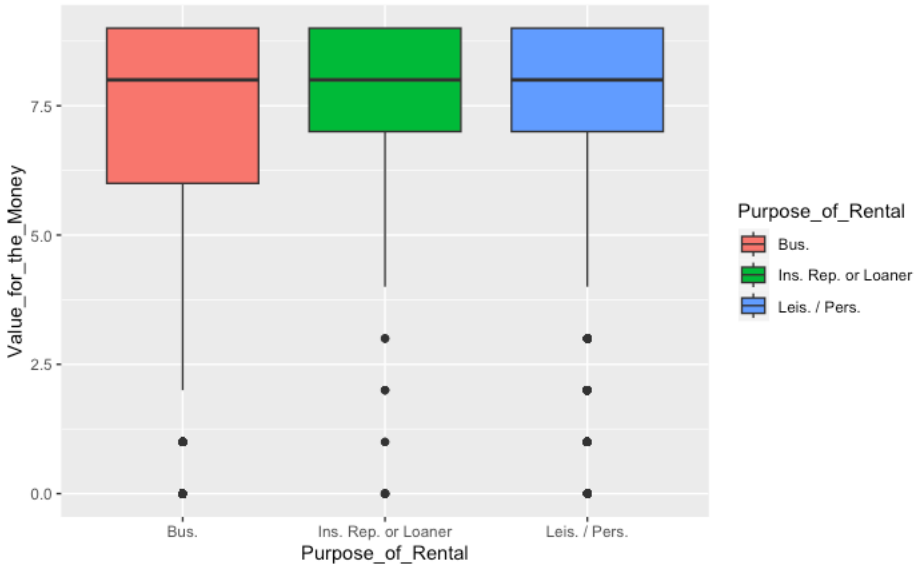


```
ggplot(siva, aes(x = charges, y = Value_for_the_Money)) +
  geom_boxplot(outlier.color = "blue", outlier.shape = 8, outlier.size = 6) +
  labs(title = "Box Plot of Value for the Money by low, average and high",
       x = "charges",
       y = "Value for the Money")
```



siva

```
ggplot(siva, aes(x=Purpose_of_Rental, y=Value_for_the_Money ,
fill=Purpose_of_Rental)) + geom_boxplot()
```

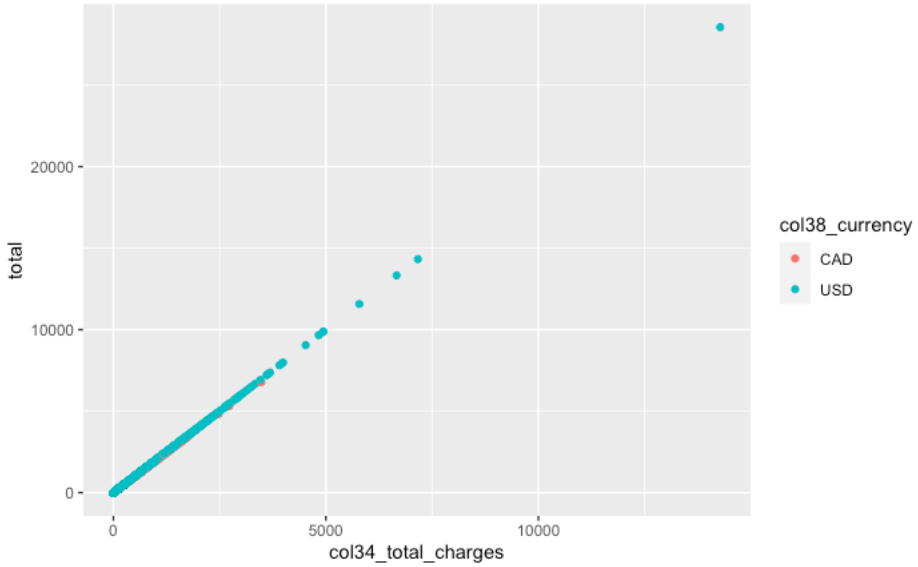


A box plot, also known as a box-and-whisker plot, provides a visual summary of the distribution of a dataset, displaying the median, quartiles, and potential outliers. It helps in understanding the spread and central tendency of the data, making it easy to identify skewness and the presence of extreme values.

The Value for money and purpose of rental boxplot shows us that overall a business rental had lower value for money than the other two categories (ins. Rep or loaner and leisure).

#5 Scatterplot

siva
library(ggplot2)# Create a scatter plot between subjects
ggplot(data = siva) + geom_point(mapping = aes(x = col34_total_charges, y =total, color = col38_currency))



A scatter plot displays individual data points as dots on a two-dimensional plane, with one variable represented on the x-axis and another on the y-axis. It is used to visualize the relationship between two continuous variables, showing patterns, trends, clusters, or the absence of any discernible relationship between the variables.

The above plot shows visually that more people paid with USD than CAD. More importantly, it shows a linear, positive relationship between total charges (a variable I created consisting of total charges in USD + total actual bill paid) and total actual bill paid. As total actual bill paid increases, so do the total charges.