|  |  |  |
| --- | --- | --- |
| A picture containing text, sign, clipart  Description automatically generated | **BOSTON**  **UNIVERSITY** | **METROPOLITAN COLLEGE** |

**AD 699 DATA MINING FOR BUSINESS ANALYTICS**

**ASSIGNMENT 1**

**FEB 17, 2023**

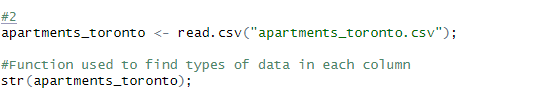
**Aravind Hanumantha Rao**

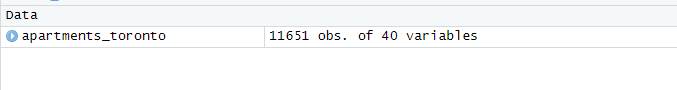
**BU ID - U55859882**

Answer 1 – Downloaded the working directory from the blackboard

**Read this file into your R environment (if it takes a while for the file to load, don’t worry -- this is normal. Be patient). Be sure to use the read.csv() function to import this dataset. a. Call the str() function on your dataset, and show the results. b. What does this function accomplish? How many rows and how many columns does your dataframe contain?**

Answer 2 –





#Function used to find types of data in each column

> str(apartments\_toronto);

'data.frame': 11651 obs. of 40 variables:

$ X\_id : int 2406105 2406106 2406107 2406108 2406109 2406110 2406111 2406112 2406113 2406114 ...

$ RSN : int 5186997 5118732 5156142 5156008 5156127 5132414 5207679 5175953 4154229 5156112 ...

$ YEAR\_REGISTERED : num NA 2022 NA 2022 2022 ...

$ YEAR\_EVALUATED : num NA NA NA NA NA NA NA NA NA NA ...

$ YEAR\_BUILT : num 2019 2021 NA 1885 1972 ...

$ PROPERTY\_TYPE : chr "PRIVATE" "PRIVATE" "PRIVATE" "PRIVATE" ...

$ WARD : int 12 13 13 9 13 11 12 10 7 12 ...

$ WARDNAME : chr "Toronto-St. Paul's" "Toronto Centre" "Toronto Centre" "Davenport" ...

$ SITE\_ADDRESS : chr "200 MADISON AVE" "25 NICHOLAS AVE" "109 PEMBROKE ST" "267 BROCK AVE" ...

$ CONFIRMED\_STOREYS : int 6 29 4 3 30 43 20 36 4 4 ...

$ CONFIRMED\_UNITS : int 82 346 20 11 240 595 177 286 144 25 ...

$ EVALUATION\_COMPLETED\_ON : chr "2023-01-31" "2023-01-27" "2023-01-27" "2023-01-24" ...

$ SCORE : int 100 100 67 73 93 99 100 100 85 75 ...

$ RESULTS\_OF\_SCORE : chr "Evaluation needs to be conducted in 3 years" "Evaluation needs to be conducted in 3 years" "Evaluation needs to be conducted in 2 years" "Evaluation needs to be conducted in 2 years" ...

$ NO\_OF\_AREAS\_EVALUATED : int 19 20 14 14 20 19 19 20 19 15 ...

$ ENTRANCE\_LOBBY : num 5 5 3 3 5 5 5 5 4 3 ...

$ ENTRANCE\_DOORS\_WINDOWS : num 5 5 4 3 5 5 5 5 4 3 ...

$ SECURITY : num 5 5 3 4 5 5 5 5 5 5 ...

$ STAIRWELLS : num 5 5 3 4 5 4 5 5 4 5 ...

$ LAUNDRY\_ROOMS : num 5 5 2 NA 5 NA NA 5 4 3 ...

$ INTERNAL\_GUARDS\_HANDRAILS : num 5 5 3 4 5 5 5 5 5 4 ...

$ GARBAGE\_CHUTE\_ROOMS : num 5 5 NA NA 4 5 5 5 4 NA ...

$ GARBAGE\_BIN\_STORAGE\_AREA : num 5 5 4 5 4 5 5 5 4 4 ...

$ ELEVATORS : num 5 5 NA 5 4 5 5 5 4 NA ...

$ STORAGE\_AREAS\_LOCKERS : num 5 5 NA NA 5 5 5 5 4 NA ...

$ INTERIOR\_WALL\_CEILING\_FLOOR: num 5 5 3 3 5 5 5 5 5 3 ...

$ INTERIOR\_LIGHTING\_LEVELS : num 5 5 3 4 4 5 5 5 5 3 ...

$ GRAFFITI : num 5 5 5 3 5 5 5 5 5 5 ...

$ EXTERIOR\_CLADDING : num 5 5 3 2 5 5 5 5 4 3 ...

$ EXTERIOR\_GROUNDS : num 5 5 3 4 5 5 5 5 4 4 ...

$ EXTERIOR\_WALKWAYS : num 5 5 4 3 5 5 5 5 4 4 ...

$ BALCONY\_GUARDS : num 5 5 NA NA 4 5 5 5 4 NA ...

$ WATER\_PEN\_EXT\_BLDG\_ELEMENTS: num 5 5 4 4 4 5 5 5 5 4 ...

$ PARKING\_AREA : num 5 5 NA NA 4 5 5 5 3 3 ...

$ OTHER\_FACILITIES : num NA 5 NA NA 5 5 5 5 NA NA ...

$ GRID : chr "S1237" "S1330" "S1328" "S0936" ...

$ LATITUDE : num 43.7 NA 43.7 43.6 43.7 ...

$ LONGITUDE : num -79.4 NA -79.4 -79.4 -79.4 ...

$ X : num 312300 316226 315042 310019 314347 ...

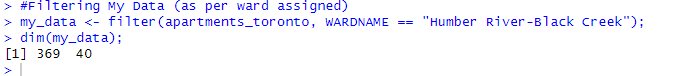
$ Y : num 4837164 4835309 4835360 4834170 4835607 ...

The read csv function , read the apartment Toronto csv file and the string apartment Toronto has a function that calculates 1156 observations of 40 variables

**3)Filter your dataset, so that it only contains records with your assigned ward (a list of all ward name assignments can be found on Blackboard). a. How many records does your dataframe contain now?**

A picture containing text

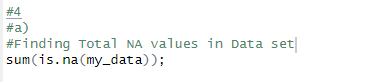
Description automatically generated

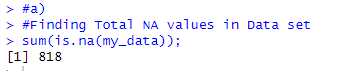


369 col 40 rows

Answer 4 –

**Are there any NA values in your dataframe? How do you know this? What is the total number of NAs in the dataframe?**

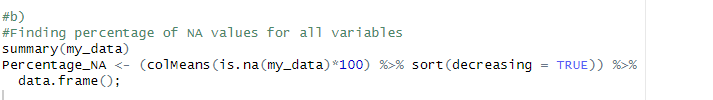
****

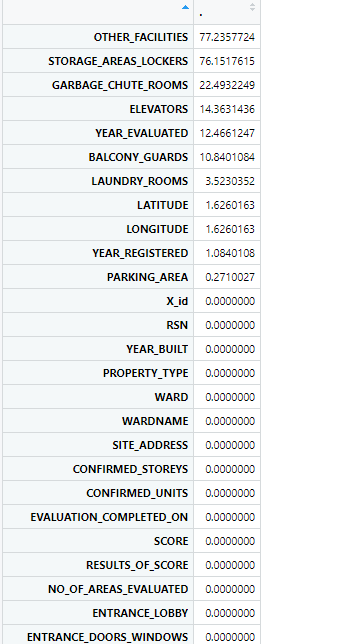


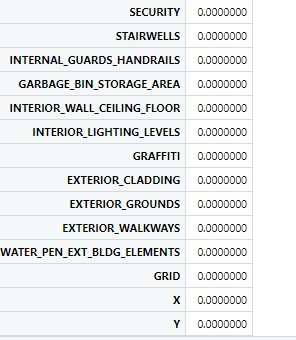
1. Yes, there are 818 NA values.

**Generate a table that shows the number of missing values and the percentage of missing values for each variable. Which variables have missing values? Pick any three columns with missing values – in a sentence or two for each one, explain why the column might have missing values. (Note: There is no domain knowledge needed to answer this question – any thoughtful explanation here is fine). Remember to consult the dataset description. You may wish to use a function from the naniar package for this step, but you are not required to.?**

1. Other facilities, storage are lockers and garbage chute rooms have the highest number of missing values. the reason for it to be having missing is due to many reasons. It can be from from data transfer or mitigation might have lost the data or else these numbers weren’t available for the certain values







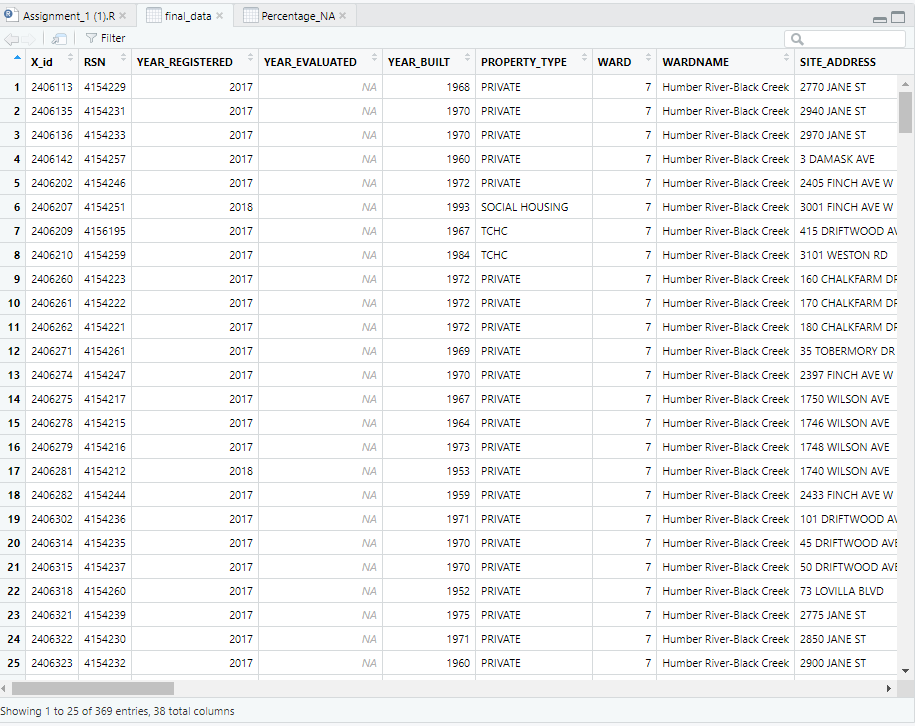


**For any column(s) whose values are more than 50% missing, remove the column(s) entirely.**

**c)**







Sample database on how it looks after removing 50 percent of the columns .

Answer 5 –

**Which column in this dataset contains dates? Run the str() function to see how R views this variable. What data type is it seen as?**

****

|  |
| --- |
| 'data.frame': 369 obs. of 38 variables:  $ X\_id : int 2406113 2406135 2406136 2406142 2406202 2406207 2406209 2406210 2406260 2406261 ...  $ RSN : int 4154229 4154231 4154233 4154257 4154246 4154251 4156195 4154259 4154223 4154222 ...  $ YEAR\_REGISTERED : num 2017 2017 2017 2017 2017 ...  $ YEAR\_EVALUATED : num NA NA NA NA NA NA NA NA NA NA ...  $ YEAR\_BUILT : num 1968 1970 1970 1960 1972 ...  $ PROPERTY\_TYPE : chr "PRIVATE" "PRIVATE" "PRIVATE" "PRIVATE" ...  $ WARD : int 7 7 7 7 7 7 7 7 7 7 ...  $ WARDNAME : chr "Humber River-Black Creek" "Humber River-Black Creek" "Humber River-Black Creek" "Humber River-Black Creek" ...  $ SITE\_ADDRESS : chr "2770 JANE ST" "2940 JANE ST" "2970 JANE ST" "3 DAMASK AVE" ...  $ CONFIRMED\_STOREYS : int 4 13 14 3 27 14 16 18 23 23 ...  $ CONFIRMED\_UNITS : int 144 153 164 10 258 166 135 176 466 262 ...  $ EVALUATION\_COMPLETED\_ON : chr "2023-01-19" "2022-12-30" "2022-12-30" "2022-12-30" ...  $ SCORE : int 85 82 86 83 75 80 81 61 80 76 ...  $ RESULTS\_OF\_SCORE : chr "Evaluation needs to be conducted in 2 years" "Evaluation needs to be conducted in 2 years" "Evaluation needs to be conducted in 3 years" "Evaluation needs to be conducted in 2 years" ...  $ NO\_OF\_AREAS\_EVALUATED : int 19 18 18 15 19 19 17 20 18 18 ...  $ ENTRANCE\_LOBBY : num 4 5 5 3 4 5 4 3 5 4 ...  $ ENTRANCE\_DOORS\_WINDOWS : num 4 4 4 5 4 4 4 3 5 5 ...  $ SECURITY : num 5 5 5 5 4 5 5 4 5 5 ...  $ STAIRWELLS : num 4 4 4 3 2 4 4 1 3 3 ...  $ LAUNDRY\_ROOMS : num 4 5 5 3 4 4 4 4 4 4 ...  $ INTERNAL\_GUARDS\_HANDRAILS : num 5 4 4 5 3 4 4 2 3 2 ...  $ GARBAGE\_CHUTE\_ROOMS : num 4 4 4 NA 3 4 NA 2 2 2 ...  $ GARBAGE\_BIN\_STORAGE\_AREA : num 4 3 4 4 4 4 3 4 4 3 ...  $ ELEVATORS : num 4 4 4 NA 5 5 4 4 4 3 ...  $ INTERIOR\_WALL\_CEILING\_FLOOR: num 5 4 4 3 3 3 4 2 4 3 ...  $ INTERIOR\_LIGHTING\_LEVELS : num 5 3 5 5 4 5 4 4 5 5 ...  $ GRAFFITI : num 5 5 5 5 5 3 5 2 5 5 ...  $ EXTERIOR\_CLADDING : num 4 4 4 4 4 4 5 4 4 4 ...  $ EXTERIOR\_GROUNDS : num 4 4 4 5 4 3 3 3 5 5 ...  $ EXTERIOR\_WALKWAYS : num 4 4 4 5 3 4 3 4 5 5 ...  $ BALCONY\_GUARDS : num 4 4 4 NA 4 4 4 3 3 4 ...  $ WATER\_PEN\_EXT\_BLDG\_ELEMENTS: num 5 4 4 3 4 4 5 4 3 3 ...  $ PARKING\_AREA : num 3 4 4 4 3 3 4 3 3 3 ...  $ GRID : chr "W0729" "W0729" "W0729" "W0728" ...  $ LATITUDE : num 43.7 43.8 43.8 43.7 43.8 ...  $ LONGITUDE : num -79.5 -79.5 -79.5 -79.5 -79.5 ...  $ X : num 303500 303352 303350 301590 301238 ...  $ Y : num 4845049 4845752 4845935 4844415 4845408 ... |
|  |
| |  | | --- | | > | |

1. The data type that had evaluation completed on has a data type of date

**Using any method, convert this variable to a ‘Date’ data type, and show that its type has been successfully converted. (Note: Be careful to pay attention to the particular way the date is written – if you don’t, this will not come out the way you want it to).**



final\_data$EVALUATION\_COMPLETED\_ON <- as.Date(final\_data$EVALUATION\_COMPLETED\_ON, format = "%Y-%m-%d");

> str(final\_data$EVALUATION\_COMPLETED\_ON)

Date[1:369], format: "2023-01-19" "2022-12-30" "2022-12-30" "2022-12-30" "2022-12-23" "2022-12-23" "2022-12-23" ...

Answer 6 –

**a)Should “Ward” be considered a numeric or categorical variable? Why?**

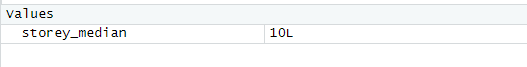
****

Since each number in the "Ward" column corresponds to a specific ward, it is appropriate to treat "Ward" as a categorical variable.

b)

**1)What is the median number of confirmed storeys for the buildings in your ward?**





**2)What is the mean number of confirmed storeys for the buildings in your ward?**

****

****

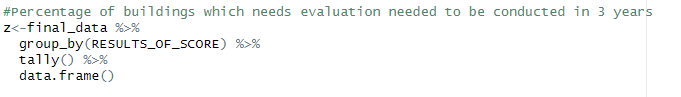
**Write a sentence or two that could help to explain the difference (or similarity) between the two average storey values that you just found. What might explain this?**

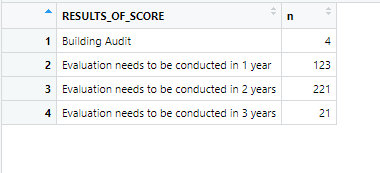
The average storey mean of 10.86 indicates the arithmetic average of all the storeys in the dataset, whereas the median storey value of 10L represents the middle value in the ordered list of storeys.

The fact that the mean and median are similar suggests that the distribution of storey values is roughly symmetric. However, if the mean is significantly higher than the median, it could suggest the presence of outliers or extreme values in the upper end of the distribution.

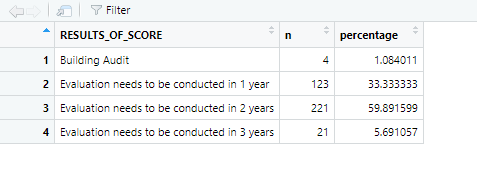
In this case, since the mean and median are relatively close, it is likely that the data is fairly normally distributed with a moderate degree of variability

**What percentage of all the buildings in your ward received a result of “Evaluation needs to be conducted in 3 years”?**

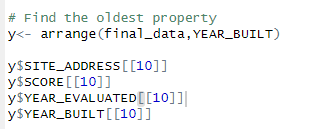
****

****

****

****

**What is the oldest building in your ward? What overall evaluation score did it earn? (Note: If several buildings are tied for ‘oldest’ because they were built in the same year, you can pick any of the ones from that year to answer this).**

****

> y$SITE\_ADDRESS[[10]]

[1] "1744 WILSON AVE"

> y$SCORE[[10]]

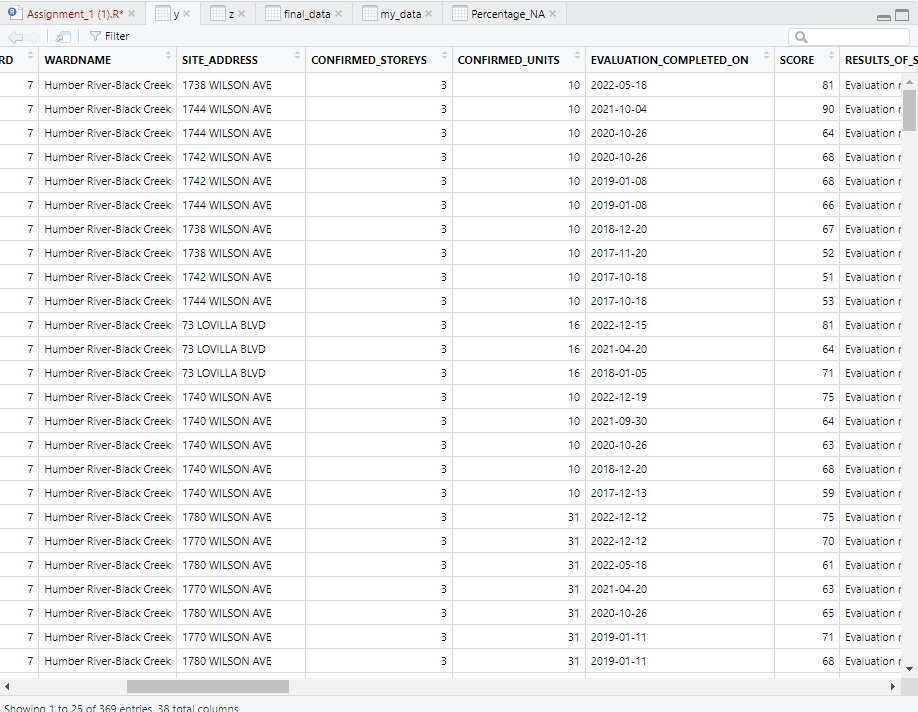
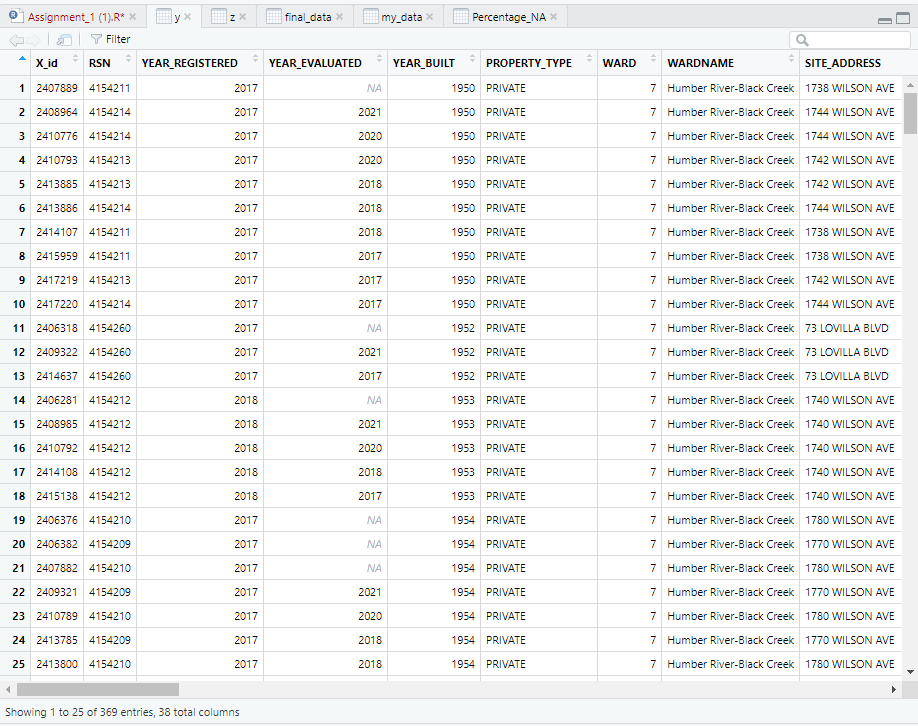
[1] 53

> y$YEAR\_EVALUATED[[10]]

[1] 2017

> y$YEAR\_BUILT[[10]]

[1] 1950

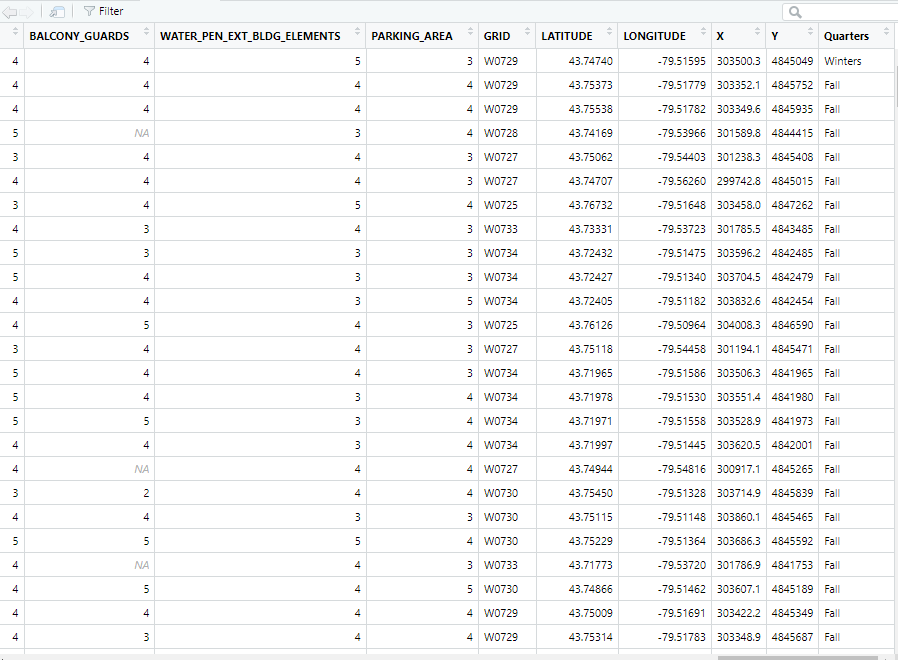
****

Answer 7-

**Using the quarter() function from lubridate, create a new column called season. Season should be created from the Evaluation\_Completed\_On variable. Next, rename the quarters so that Quarter 1 becomes “Winter”, Quarter 2 becomes “Spring”, Quarter 3 becomes “Summer” and Quarter 4 becomes “Fall.”**

A picture containing text

Description automatically generated





> final\_1$Quarters

[1] "Winters" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall"

[11] "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall"

[21] "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall"

[31] "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall"

[41] "Spring" "Spring" "Spring" "Spring" "Spring" "Spring" "Fall" "Fall" "Fall" "Fall"

[51] "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall"

[61] "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall"

[71] "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall"

[81] "Fall" "Fall" "Fall" "Summer" "Summer" "Summer" "Spring" "Spring" "Spring" "Spring"

[91] "Spring" "Spring" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall"

[101] "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall"

[111] "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall"

[121] "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall"

[131] "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall"

[141] "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall"

[151] "Fall" "Fall" "Fall" "Fall" "Summer" "Winters" "Fall" "Fall" "Fall" "Fall"

[161] "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall"

[171] "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall"

[181] "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall"

[191] "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall"

[201] "Spring" "Winters" "Winters" "Winters" "Winters" "Winters" "Winters" "Winters" "Winters" "Winters"

[211] "Winters" "Winters" "Winters" "Winters" "Winters" "Winters" "Winters" "Winters" "Winters" "Winters"

[221] "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall"

[231] "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall"

[241] "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall"

[251] "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall"

[261] "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Winters"

[271] "Winters" "Winters" "Winters" "Winters" "Winters" "Winters" "Winters" "Winters" "Winters" "Winters"

[281] "Winters" "Winters" "Winters" "Winters" "Winters" "Winters" "Winters" "Fall" "Fall" "Fall"

[291] "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall"

[301] "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall"

[311] "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall"

[321] "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall"

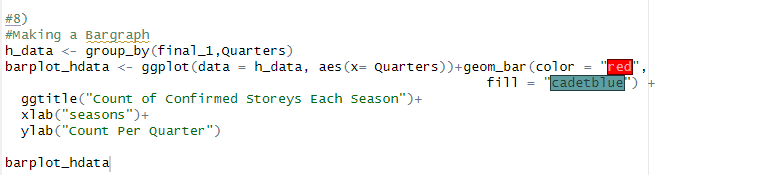
[331] "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall"

[341] "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall"

[351] "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall"

[361] "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Fall" "Summer" "Summer"

**Using ggplot, construct a barplot showing the counts of completed evaluations during each of the four seasons. Fill your bars with any color of your choice. a. What do you notice about your plot? Why might it look the way it does? (note: there is NO need for domain knowledge here – just take a moment to think about it, and answer with any reasonable speculation on your part)?**

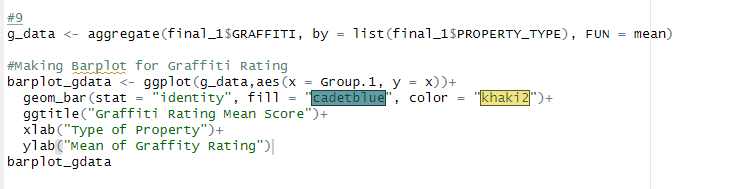


Chart, bar chart

Description automatically generated

During the fall season, the real estate market saw a higher volume of property evaluations compared to other seasons, which experienced lower levels of activity. This trend suggests that fall is a popular time for property owners and potential buyers to assess the value of properties, potentially influenced by factors such as the upcoming tax season or favorable weather conditions for viewing properties. Understanding seasonal trends in property evaluations can help real estate professionals better anticipate market activity and tailor their services accordingly.

**Again using ggplot, let’s make another barplot. This time, place the property types on the x-axis. On the y-axis, show the mean scores for graffiti ratings for each property type. a. What does this plot show you? Write 1-2 sentences that speculate about potential reasons for the different ratings among the property types. (remember that 5 is the ‘best’, or cleanest, graffiti score)?**

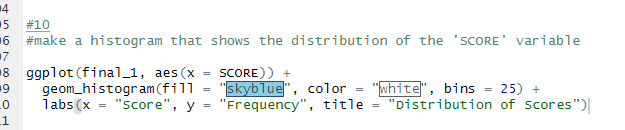
****

Chart, bar chart

Description automatically generated

private housing appears to have the best mean graffiti rating compared to other building types, with a rating of 5. It is possible that private housing owners invest more in graffiti prevention measures, such as surveillance cameras, security guards, or protective coatings on building surfaces. Alternatively, private housing residents may be more vigilant in reporting graffiti and working with local authorities to promptly remove it. Further investigation into the specific factors that contribute to the different graffiti ratings among property types could provide valuable insights for property owners, community organizations, and local governments in addressing urban blight and promoting neighborhood revitalization.

**10)Using ggplot, make a histogram that shows the distribution of the ‘SCORE’ variable. Use any number of bins, and stylize it with any color/fill values of your choice**.



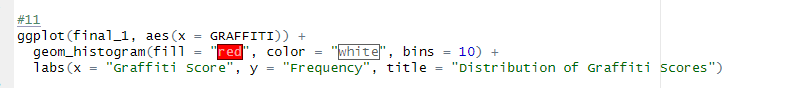
Chart, histogram

Description automatically generated

**In a sentence or two, describe this histogram. What does it show about the scores?**

The histogram of evaluation scores suggests that the distribution of scores is approximately normal, with a mean score of 70. This indicates that the majority of properties are receiving a moderately good evaluation score, with relatively fewer properties receiving very low or very high scores. Additionally, the requirement to evaluate every two years suggests a regular and systematic process for assessing the value of real estate properties, which can help ensure transparency and accountability in the real estate market.

**11)Using ggplot, make a histogram that shows the distribution of the ‘GRAFFITI’ variable. b. In a sentence or two, describe your plot – what does it show?**



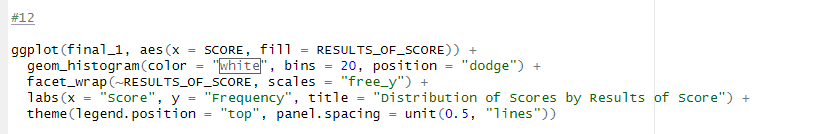
Chart, bar chart

Description automatically generated

**In a sentence or two, describe your plot – what does it show?**

The plot of graffiti scores indicates that a rating of 5 has the highest frequency of occurrence compared to other scores, while a rating of 2 has the lowest frequency. This suggests that properties with a rating of 5 have the lowest incidence of graffiti, while properties with a rating of 2 are more likely to have visible graffiti.

**12)Now, generate faceted histograms. These histograms should depict the distribution of the SCORE variable, faceted on the RESULTS\_OF\_SCORE variable. Fill your histograms with any color of your choice, and use whatever number of bins that you wish to use**



Chart

Description automatically generated

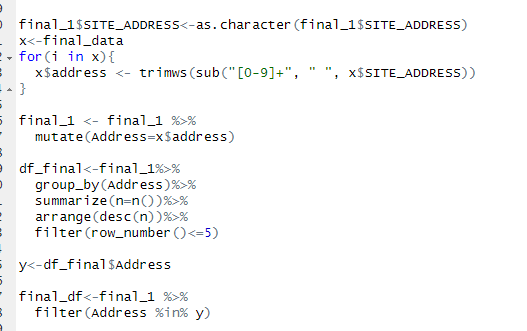
**What do you see here? In a few sentences, describe what these faceted histograms show. What connection can you make between the score values and the ‘results of score’ outcomes?**

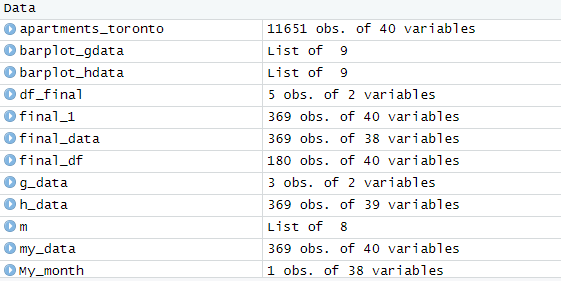
The faceted histograms provide a visual representation of how the number of years between property evaluations is related to the evaluation scores. The plots show that, generally, properties that require evaluation over longer intervals tend to have higher scores. Conversely, properties with shorter evaluation intervals tend to have lower scores. The histograms also reveal that the distribution of evaluation scores changes in shape and skewness depending on the evaluation interval.

Question 13-

**Okay, so it’s time for one more filter operation. Filter the dataframe so that only the five most common streets from your ward remain. You can approach this any way that you would like to – but you may wish to first split the SITE\_ADDRESS variable into separate columns to make this easier?**

1. **Filter your dataframe so that it only contains properties from the five most common streets in your ward.**





Graphical user interface

Description automatically generated

> final\_df$Address

[1] "JANE ST" "JANE ST" "JANE ST" "FINCH AVE W" "FINCH AVE W" "WESTON RD"

[7] "CHALKFARM DR" "CHALKFARM DR" "CHALKFARM DR" "FINCH AVE W" "WILSON AVE" "WILSON AVE"

[13] "WILSON AVE" "WILSON AVE" "FINCH AVE W" "JANE ST" "JANE ST" "JANE ST"

[19] "WILSON AVE" "WILSON AVE" "WILSON AVE" "JANE ST" "FINCH AVE W" "WILSON AVE"

[25] "WILSON AVE" "WESTON RD" "FINCH AVE W" "FINCH AVE W" "JANE ST" "JANE ST"

[31] "JANE ST" "WESTON RD" "WESTON RD" "JANE ST" "JANE ST" "CHALKFARM DR"

[37] "WILSON AVE" "FINCH AVE W" "JANE ST" "WESTON RD" "WESTON RD" "WILSON AVE"

[43] "WILSON AVE" "WESTON RD" "WESTON RD" "WILSON AVE" "WILSON AVE" "JANE ST"

[49] "JANE ST" "JANE ST" "JANE ST" "JANE ST" "FINCH AVE W" "FINCH AVE W"

[55] "FINCH AVE W" "JANE ST" "FINCH AVE W" "FINCH AVE W" "JANE ST" "WESTON RD"

[61] "WESTON RD" "WILSON AVE" "WILSON AVE" "WILSON AVE" "WILSON AVE" "WILSON AVE"

[67] "WILSON AVE" "CHALKFARM DR" "CHALKFARM DR" "CHALKFARM DR" "JANE ST" "JANE ST"

[73] "JANE ST" "JANE ST" "WESTON RD" "FINCH AVE W" "JANE ST" "JANE ST"

[79] "FINCH AVE W" "WESTON RD" "WESTON RD" "WESTON RD" "WESTON RD" "WESTON RD"

[85] "JANE ST" "JANE ST" "JANE ST" "JANE ST" "JANE ST" "JANE ST"

[91] "FINCH AVE W" "FINCH AVE W" "FINCH AVE W" "WILSON AVE" "CHALKFARM DR" "CHALKFARM DR"

[97] "CHALKFARM DR" "WESTON RD" "WESTON RD" "FINCH AVE W" "FINCH AVE W" "JANE ST"

[103] "JANE ST" "WILSON AVE" "WILSON AVE" "WILSON AVE" "WILSON AVE" "WILSON AVE"

[109] "WILSON AVE" "WILSON AVE" "WILSON AVE" "JANE ST" "JANE ST" "WILSON AVE"

[115] "WILSON AVE" "CHALKFARM DR" "CHALKFARM DR" "CHALKFARM DR" "CHALKFARM DR" "FINCH AVE W"

[121] "JANE ST" "JANE ST" "JANE ST" "JANE ST" "JANE ST" "WESTON RD"

[127] "FINCH AVE W" "FINCH AVE W" "WESTON RD" "JANE ST" "JANE ST" "FINCH AVE W"

[133] "FINCH AVE W" "FINCH AVE W" "JANE ST" "JANE ST" "FINCH AVE W" "WILSON AVE"

[139] "WESTON RD" "WILSON AVE" "FINCH AVE W" "FINCH AVE W" "JANE ST" "FINCH AVE W"

[145] "WESTON RD" "JANE ST" "JANE ST" "JANE ST" "JANE ST" "WILSON AVE"

[151] "WILSON AVE" "WILSON AVE" "WILSON AVE" "WILSON AVE" "WILSON AVE" "CHALKFARM DR"

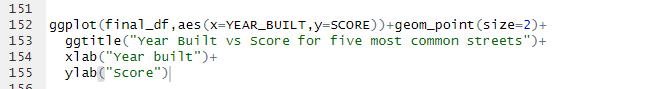
[157] "JANE ST" "JANE ST" "JANE ST" "JANE ST" "FINCH AVE W" "CHALKFARM DR"

[163] "CHALKFARM DR" "CHALKFARM DR" "FINCH AVE W" "WILSON AVE" "WILSON AVE" "WESTON RD"

[169] "WESTON RD" "FINCH AVE W" "FINCH AVE W" "WESTON RD" "WESTON RD" "WESTON RD"

[175] "WESTON RD" "JANE ST" "JANE ST" "JANE ST" "JANE ST" "JANE ST"

1. **Now, build a scatterplot with this newly-filtered version of your data. Place YEAR\_BUILT along your x-axis. Place SCORE on your y-axis, and use different colors for each street name. What do you see here? Are there any patterns or interesting takeaways from this graph?**

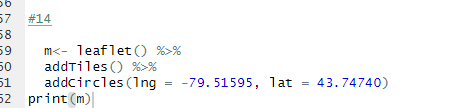
****

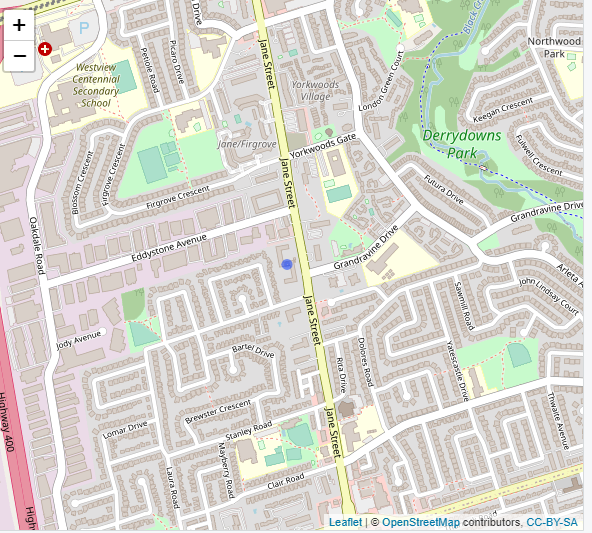
Chart, scatter chart

Description automatically generated

The graph shows the top 5 most common streets in the dataset and their corresponding number of properties. The data has been shortened to include only the top 5 streets, with JANE\_ST having the highest number of properties at 59 and CHALKFARM DR having the lowest at 18. Additionally, the graph shows that the three clusters of 1950, 1960, and 1970 have the highest concentration of properties, while later years have fewer properties. This suggests that the majority of properties in the dataset were constructed during these decades, which may reflect historical trends in real estate development and urbanization. Understanding the distribution of properties by street and construction date can be useful for real estate professionals and property owners to identify potential opportunities and challenges in different neighborhoods and time periods.

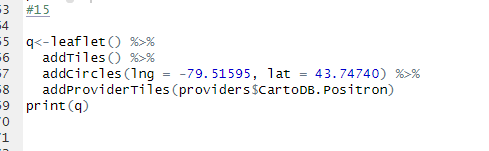
m <- leaflet() %>% addTiles() %>% addCircles(lng= ? , lat= ?) m # Print the map





Run something similar, but this time, select something of your choice after the dollar sign. If you’re not sure what to choose, try a few things out and explore to find out what they do!

m <- leaflet() %>% addTiles() %>% addCircles(lng= ? , lat= ? ) %>% addProviderTiles(providers$\_\_\_\_\_\_\_\_\_\_\_\_\_) m # Print the map



Diagram

Description automatically generated