Project-Group-6-Proposal

October 29, 2021

1 Final Project Proposal - Group 6

1.0.1 Group Members:

Hamza Hassan (100788913)

Shriraam Murugathas (100622836)

Imran Mustafa (100786010)

Raj Shukla (100784045)

1.1 Getting The Data

1.1.1 Description:

This data relates to the RentSafeTO program in Toronto, it contains evaluations of apartment buildings with three or more storeys or 10 or more units. Each evaluation (conducted by an inspector) ranks various aspects of the complex, mechanical and security systems, parking and exterior grounds, ect. They are ranked from 1-5 with five being the best score. This score helps determine the building's condition and this information is then used to help decision making about the complex.

an overview of the data as well as where it was gotten from, is available at https://open.toronto.ca/dataset/apartment-building-evaluation/. The downloaded copy is also in the same dir as the notebook.

Note: Due to the data being updated daily, it is important to know this copy was created on October 22, 2021, 17:53:55

[14]:
```
import pandas as pd

# code to port the data into a pandas dataframe
with open('Apartment Building Evaluation.csv', 'r') as f:
    df = pd.read_csv(f)

# setting some new defaults
df.style.hide_index()
pd.set_option("display.max.columns", None)
%matplotlib inline

# simple output statement to see if all of the code works well
```

```
df.head()
```

[14]:

	_id	RSN	YEAR_REGISTERED	YEAR_EVALUATED	YEAR_BUILT	PROPERTY_TYPE
0	429048	4155178	2017.0	2021	1960.0	PRIVATE
1	429049	4155132	2017.0	2021	1971.0	PRIVATE
2	429050	4154929	2017.0	2021	1960.0	PRIVATE
3	429051	4155950	2018.0	2021	1953.0	PRIVATE
4	429052	4153722	2017.0	2021	1962.0	TCHC

	WARD	WARDNAME	SITE_ADDRESS	CONFIRMED_STOREYS
0	5	York South-Weston	220 WOOLNER AVE	9
1	5	York South-Weston	65 EMMETT AVE	24
2	12	Toronto-St. Paul's	450 WALMER RD	15
3	19	Beaches-East York	5 STAG HILL DR	4
4	19	Beaches-East York	828 KINGSTON RD	7

	CONFIRMED_UNITS	EVALUATION_COMPLETED_ON	SCORE
0	130	2021-10-20	81
1	419	2021-10-20	79
2	171	2021-10-20	74
3	15	2021-10-20	85
4	147	2021-10-20	87

	RESULTS_OF_SCORE	NO_OF_AREAS_EVALUATED
0	Evaluation needs to be conducted in 2 years	17
1	Evaluation needs to be conducted in 2 years	19
2	Evaluation needs to be conducted in 2 years	19
3	Evaluation needs to be conducted in 2 years	15
4	Evaluation needs to be conducted in 3 years	19

	ENTRANCE_LOBBY	ENTRANCE_DOORS_WINDOWS	SECURITY	STAIRWELLS
0	4.0	4.0	5.0	3.0
1	4.0	4.0	5.0	2.0
2	3.0	4.0	5.0	3.0
3	4.0	3.0	5.0	5.0
4	3.0	3.0	5.0	5.0

	LAUNDRY_ROOMS	INTERNAL_GUARDS_HANDRAILS	GARBAGE_CHUTE_ROOMS
0	3.0	4.0	NaN
1	4.0	4.0	3.0
2	3.0	3.0	3.0
3	5.0	5.0	NaN
4	4.0	5.0	5.0

	GARBAGE_BIN_STORAGE_AREA	ELEVATORS	STORAGE_AREAS_LOCKERS
0	3.0	5.0	NaN
1	4.0	5.0	3.0

```
2                      3.0        4.0                    4.0
3                      4.0        NaN                    NaN
4                      5.0        5.0                    NaN

   INTERIOR_WALL_CEILING_FLOOR  INTERIOR_LIGHTING_LEVELS  GRAFFITI  \
0                          4.0                       5.0       5.0
1                          3.0                       4.0       5.0
2                          4.0                       3.0       5.0
3                          4.0                       3.0       5.0
4                          3.0                       5.0       5.0

   EXTERIOR_CLADDING  EXTERIOR_GROUNDS  EXTERIOR_WALKWAYS  BALCONY_GUARDS  \
0                4.0               4.0                4.0             4.0
1                4.0               4.0                5.0             3.0
2                4.0               4.0                4.0             4.0
3                3.0               5.0                5.0             NaN
4                3.0               5.0                3.0             5.0

   WATER_PEN_EXT_BLDG_ELEMENTS  PARKING_AREA  OTHER_FACILITIES   GRID  \
0                          4.0           4.0               NaN  W0539
1                          5.0           4.0               NaN  W0531
2                          4.0           3.0               NaN  S1230
3                          3.0           5.0               NaN  S1927
4                          4.0           5.0               5.0  S1936

    LATITUDE  LONGITUDE           X            Y
0  43.672378 -79.492949  305352.953  4836714.002
1  43.688042 -79.504080  304455.515  4838454.156
2  43.686218 -79.413643  311746.530  4838255.111
3  43.703302 -79.310988  319989.443  4840181.689
4  43.680524 -79.292114  321545.770  4837639.768
```

1.2 Attributes:

Each of the attributes in the data set and what they mean.

_id Unique row identifier for Open Data database

RSN This is the ID number for a building. The registration ID can be used to identify information pertaining to buildings in different RentSafeTO Open Data Sets.

YEAR_REGISTERED This is the year that the building first registered with RentSafeTO.

YEAR_EVALUATED This represents the year of the building evaluation scores.

YEAR_BUILT This is the year that the building was built in. Information is provided by the Building Owners/Managers.

PROPERTY_TYPE This field informs users of whether a building is owed privately, by Toronto Community Housing Corporation (TCHC) or another assisted, social or supportive housing provider.

WARD This is the ward that the building is located in. All data is provided based on the 25 ward system.

WARDNAME This is the name of the ward

SITE_ADDRESS This is the building address.

CONFIRMED_STOREYS This is the number of storeys in a building.

CONFIRMED_UNITS This is the number of units in a building.

EVALUATION_COMPLETED_ON This is the date the evaluation was conducted on.

SCORE This is the overall score of the building. The score is the sum total of each item that was evaluated. The formula to calculate scores is as follows: sum of all assigned scores during the evaluation / (number of unique items reviewed *5) RESULTS_OF_SCORE

The score is used to determine whether an audit (which is a comprehensive examination of the building) takes place or whether another evaluation will be conducted in one, two or three years.

NO_OF_AREAS_EVALUATED This is the number of items that were evaluated during a single evaluation.

ENTRANCE_LOBBY This represents the condition of the entrance and/or lobby in a building. Scores range from 1 to 5, with 1 being the lowest and 5 being the highest.

ENTRANCE_DOORS_WINDOWS This represents the condition of the entrance doors and windows in a building. Scores range from 1 to 5, with 1 being the lowest and 5 being the highest.

SECURITY This represents the condition of the security system(s) in a building. Scores range from 1 to 5, with 1 being the lowest and 5 being the highest.

STAIRWELLS This represents the condition of the stairwells in a building. Scores range from 1 to 5, with 1 being the lowest and 5 being the highest.

LAUNDRY_ROOMS This represents the condition of the laundry room(s) in a building. Scores range from 1 to 5, with 1 being the lowest and 5 being the highest.

INTERNAL_GUARDS_HANDRAILS This represents the condition of the internal guards and handrails in a building. Scores range from 1 to 5, with 1 being the lowest and 5 being the highest.

GARBAGE_CHUTE_ROOMS This represents the condition of the garbage/chute rooms in a building. Scores range from 1 to 5, with 1 being the lowest and 5 being the highest.

GARBAGE_BIN_STORAGE_AREA This represents the condition of the garbage bin storage room or outdoor enclosure area. Scores range from 1 to 5, with 1 being the lowest and 5 being the highest.

ELEVATORS This represents the condition of the elevator(s) in a building. Scores range from 1 to 5, with 1 being the lowest and 5 being the highest.

STORAGE_AREAS_LOCKERS This represents the condition of the storage areas/lockers in a building. Scores range from 1 to 5, with 1 being the lowest and 5 being the highest.

INTERIOR_WALL_CEILING_FLOOR This represents the condition of internal walls, ceilings and floors in a building. Scores range from 1 to 5, with 1 being the lowest and 5 being the highest.

INTERIOR_LIGHTING_LEVELS This represents the condition of internal lighting levels in a building. Scores range from 1 to 5, with 1 being the lowest and 5 being the highest.

GRAFFITI This score represents the severity of graffiti in a building. Scores range from 1 to 5, with 1 being a significant amount of graffiti and 5 being no graffiti.

EXTERIOR_CLADDING This represents the condition of the exterior cladding/bricks/paint, flashing and drain pipes on a building. Scores range from 1 to 5, with 1 being the lowest and 5 being the highest.

EXTERIOR_GROUNDS This represents the condition of the exterior grounds of a building. Scores range from 1 to 5, with 1 being the lowest and 5 being the highest.

EXTERIOR_WALKWAYS This represents the condition of the exterior walkways of a building. Scores range from 1 to 5, with 1 being the lowest and 5 being the highest.

BALCONY_GUARDS This represents the condition of the balcony guards on a building. Scores range from 1 to 5, with 1 being the lowest and 5 being the highest.

WATER_PEN_EXT_BLDG_ELEMENTS This represents the condition of water penetration of external elements of a building. Scores range from 1 to 5, with 1 being the lowest and 5 being the highest.

PARKING_AREA This represents the condition of the parking areas of a building. Scores range from 1 to 5, with 1 being the lowest and 5 being the highest.

OTHER_FACILITIES This represents the condition of other facilities in a building. Scores range from 1 to 5, with 1 being the lowest and 5 being the highest.

GRID This is the grid that the building is located in. A grid represents a specific administrative area for bylaw enforcement.

LATITUDE The latitude associated with the building address.

LONGITUDE The longitude associated with the building address.

X The projected X coordinate associated with the building address. The projected coordinate system is NAD27 MTM Zone 10.

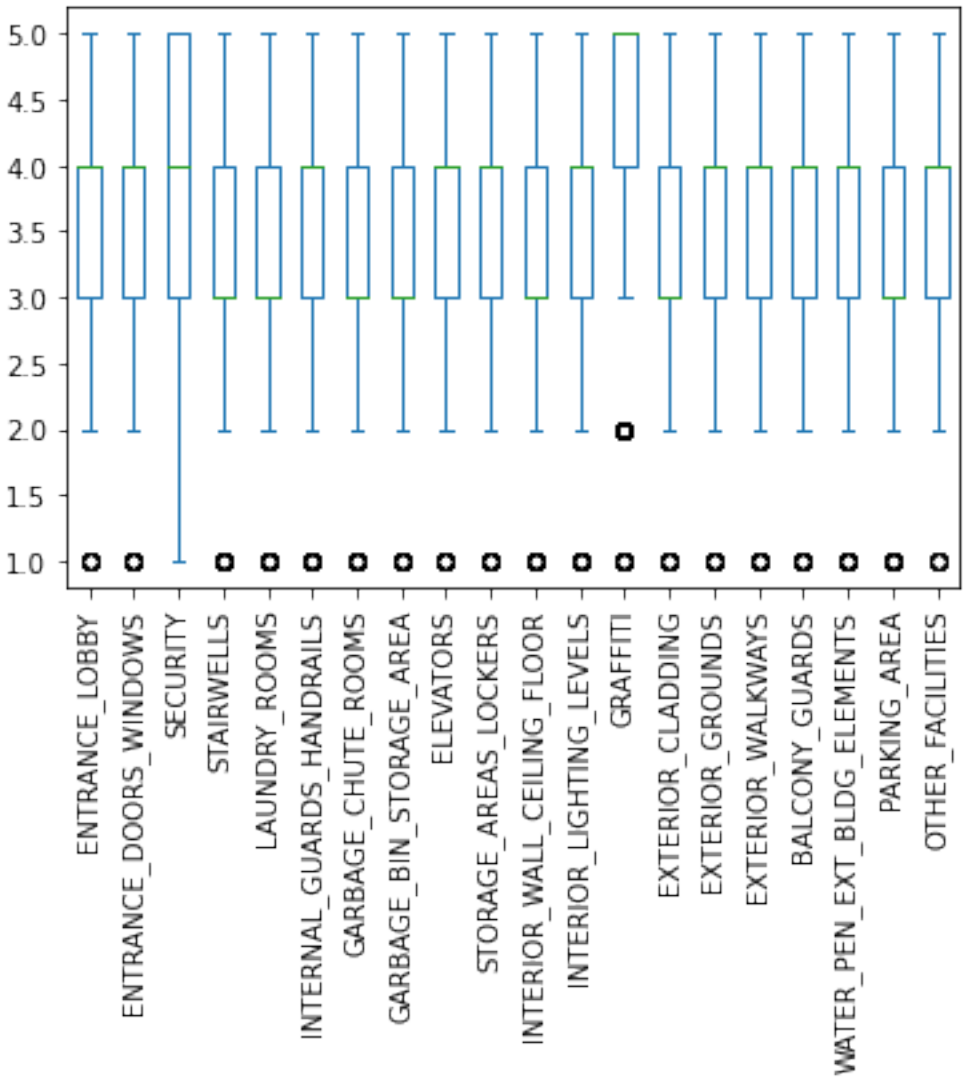
Y The projected Y coordinate associated with the building address. The projected coordinate system is NAD27 MTM Zone 10.

1.3 Getting to Know the Data

This is just some basic analysis of the data to get a better understanding of what is in it.

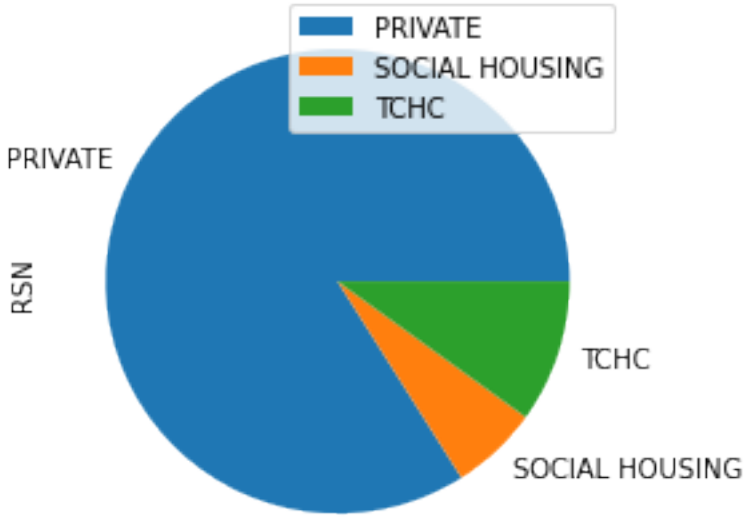
[17]:
```
# some imprtent colums put into list to make indexing easy.
attr = [
'ENTRANCE_LOBBY',
'ENTRANCE_DOORS_WINDOWS',
'SECURITY',
'STAIRWELLS',
'LAUNDRY_ROOMS',
'INTERNAL_GUARDS_HANDRAILS',
'GARBAGE_CHUTE_ROOMS',
'GARBAGE_BIN_STORAGE_AREA',
'ELEVATORS',
'STORAGE_AREAS_LOCKERS',
'INTERIOR_WALL_CEILING_FLOOR',
'INTERIOR_LIGHTING_LEVELS',
'GRAFFITI',
'EXTERIOR_CLADDING',
'EXTERIOR_GROUNDS',
'EXTERIOR_WALKWAYS',
'BALCONY_GUARDS',
'WATER_PEN_EXT_BLDG_ELEMENTS',
'PARKING_AREA',
'OTHER_FACILITIES']

#looking at the boxplot for the metrics
fig1 = df[attr].plot.box(rot=90)
print(f"{df['SCORE'].mean()} is the mean score of the data, with a min score of␣
 ↪{df['SCORE'].min()} and a max of {df['SCORE'].max()}")
print(f"form a total of {df.shape[0]} evaulations. from {df['RSN'].nunique()}␣
 ↪diffrent buildings\nThe graph shows that score borken down by each metric.")
```

71.88075313807532 is the mean score of the data, with a min score of 0 and a max of 100
form a total of 9082 evaulations. from 3476 diffrent buildings
The graph shows that score borken down by each metric.



[3]:
```
print(f"Their are a total of {df['PROPERTY_TYPE'].nunique()} Property types. Broken down they are\n{df['PROPERTY_TYPE'].value_counts()}")
fig2 = df.groupby(['PROPERTY_TYPE']).sum().plot(kind='pie', y='RSN')
```

```
Their are a total of 3 Property types. Broken down they are
PRIVATE           7625
TCHC               904
SOCIAL HOUSING     553
Name: PROPERTY_TYPE, dtype: int64
```



[15]:
```
print(f"All of the building were built between {df['YEAR_BUILT'].min()} and␣
 ↪{df['YEAR_BUILT'].max()} with the median age being {df['YEAR_BUILT'].
 ↪median()}")
print(f"With most of them having been built in ward {df['WARD'].
 ↪mode(dropna=True)}")
```

```
All of the building were built between 1805.0 and 2021.0 with the median age
being 1961.0
With most of them having been built in ward 0    12
dtype: int64
```

1.4 Justification

This data was chosen above all others as it relates to the modern issue of housing. As many Canadians are struggling to find adequate shelter, there is the other issue of the quality of the housing that many people find themselves in. whether it has the appropriate accommodations or is safe to live in. These questions are answered in this data set as the over 9000 evaluations help paint a picture of the state of housing in the Toronto area.

while there were other data sets that were interesting to the group. as a short list.

https://www.kaggle.com/uciml/student-alcohol-consumption

https://www.kaggle.com/jessicali9530/animal-crossing-new-horizons-nookplaza-dataset

https://www.kaggle.com/spscientist/students-performance-in-exams

Each of these data sets had its own reasons for analysis but the rentsafe data set came out on top in the end.

1.5 Preliminary Questions

In this proposal we will be asking some basic questions and then analyzing the data to see if these questions produce any interesting trends or insights into the data.

1.5.1 Q1 What wards have the highest/lowest scores

Seeing as the city is broken down into many wards it is a simple question to see whether any wards have overall higher scores, or score higher in particular categories, or have no effect on the final score.

this type of question will be answered by using pandas inbuilt analytical functions, to start with the groupby() function in pandas will be instrumental in this analysis as it will allow us to put all buildings into their ward and then calculate their data individually so that each group can be compared to it peers.

the standard data points will be calculated with each group, such as mean, median, mode, quartiles, and deviation These will allow some basic comparison between each ward.

after theis cov() and corr() will be used with groupby() to see if any one of the wards relates to better overall scores or overall scores in a metric. These numbers may also be graphed with pandas graphing capabilities if those can be used appropriately.

These calculations will be used to determine what if any affect the ward has on the final score

1.5.2 Q2 How does the age of the complex affect it's score

this questions deals with the age of a property and whether that will have any effects on metrics or the overall score

Using the groupby() function a mean, min, max, quartiles, as well as deviation will be calculated to compare the groups; these data points will be graphed as a boxplot with year as x and score as y.

Then after this corr() and cov() will be used simerly to determine whether a strong relationship between these data points can be found; this will be placed on a line graph to help better understand the data.

from all of this it will be determined whether the age has any affect.

1.5.3 Q3 How does the property type/ownership affect the scoring of the building

This question will have the data points broken into groups based on the type of ownership that the complex is under, each of the following groups will be run through the same calculation. The results will be compared to each other using graphs and tables to find out if any one of the ownership types affect the score in any meaningful way.

1.5.4 Q4 What are the most common attributes of Toronto apartments

An important thing to take note of this data set is that when a metric is not present in a building it will be treated as null. And this brings up the question of what are the most common features

present in Toronto housing. do all residents have access to what many would consider basic amenities in their housing elevators, laundry rooms, garbage shoots etc. This is what we seek to find out in this question.

to answer this each buildings metrics will be added together into a total from this a graph shall be made displaying the prevalence of each of these amenities. from this data it will be determined whether there are any gaps in the amenities that are provided in Toronto housing. which will be summarized in the report

1.5.5 Q5 Does Security affect the level of graffiti on the complex

This simple question seeks to answer whether the money many of these buildings are spending on security (CCTv, garuds, locks etc.) correlates with any level of change with the amount of graffiti present on the complex.

To answer this question, the relationship will be graphed to see the trend of security (x-axis) and graffiti (y-axis). then a table containing the correlation seen between these two data points will be made, and the results summarized in the final report.

1.6 Further Application

The insights gained from the data can be used to see what modern Toronto housing is like, and most importantly what areas it needs to improve in, this will allow officials to better allocate funding and programs to areas that both need improvement and are of importance to the city and its residents. For example if there is any correlation between property type and the rating of the building, then this might lead to that type of property receiving more funding/regulation. And when the data is used like this it will be invaluable in shaping the future of the city.