CMPE-255

Jaswanth Karangula
jaswanth.karangula@sjsu.edu

Dataset:

For this assignment i have selected the House price prediction data set from Kaggle
Data set link : https://www.kaggle.com/datasets/shree1992/housedata

I have uploaded this dataset using the google cloud console UI
The uploading procedure has been recorded and can be found here
https://drive.google.com/file/d/1uKGlgT2bf9RSHkPhwvVgA2qOmmPmlr7K/view?usp=sharing

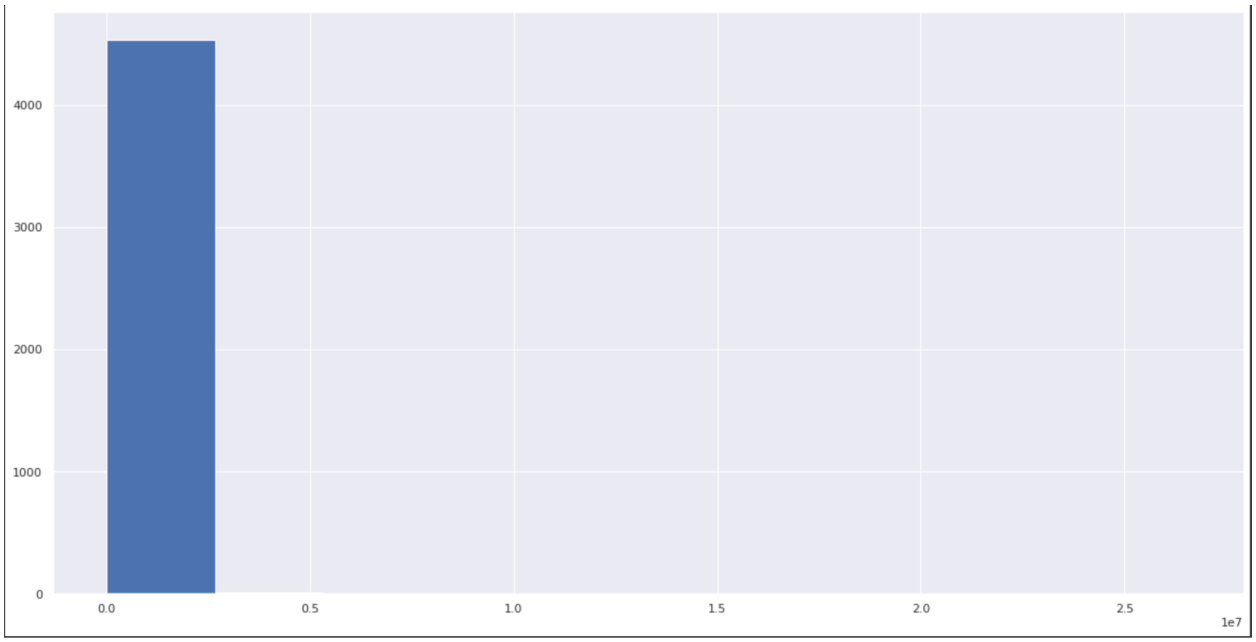
Code:
Colab Notebook Link :
https://colab.research.google.com/drive/1CzUylxm43KLbTiNxfLbLv0uhlCh3HIDB#scrollTo=wuCj5q3EdEuV

I have performed EDA on the House Price data and tried to visualize some graphs

Firstly I have loaded the dataset from bigquery into the colab and used pandas and seaborn to visualize the the dataframes

Graph 1:

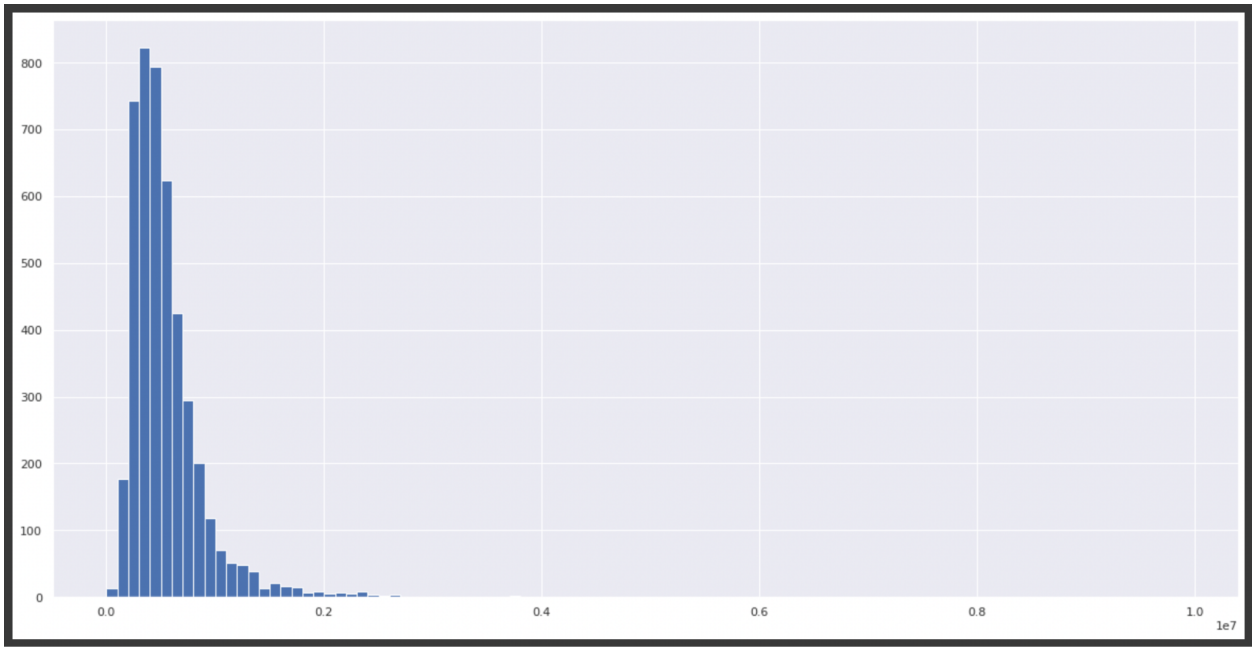
Drawn a sample hist graph for the house prices



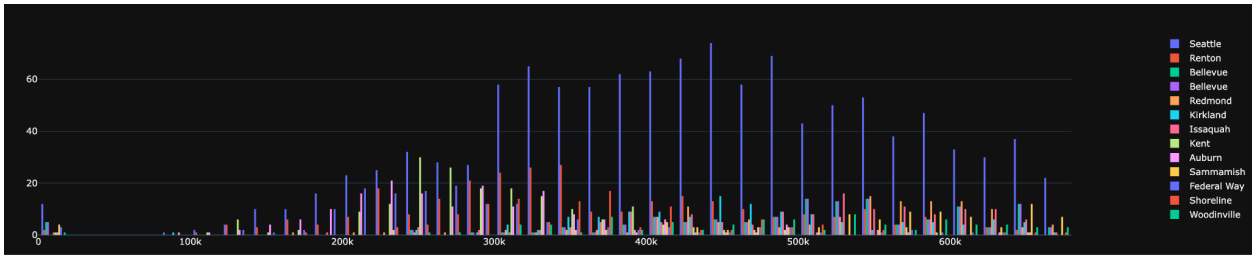
Then after adjusting the bins the prices were binned perfectly into different bins

Used the bins as

```
price_bins = range(1000, 10000000, 100000)
```

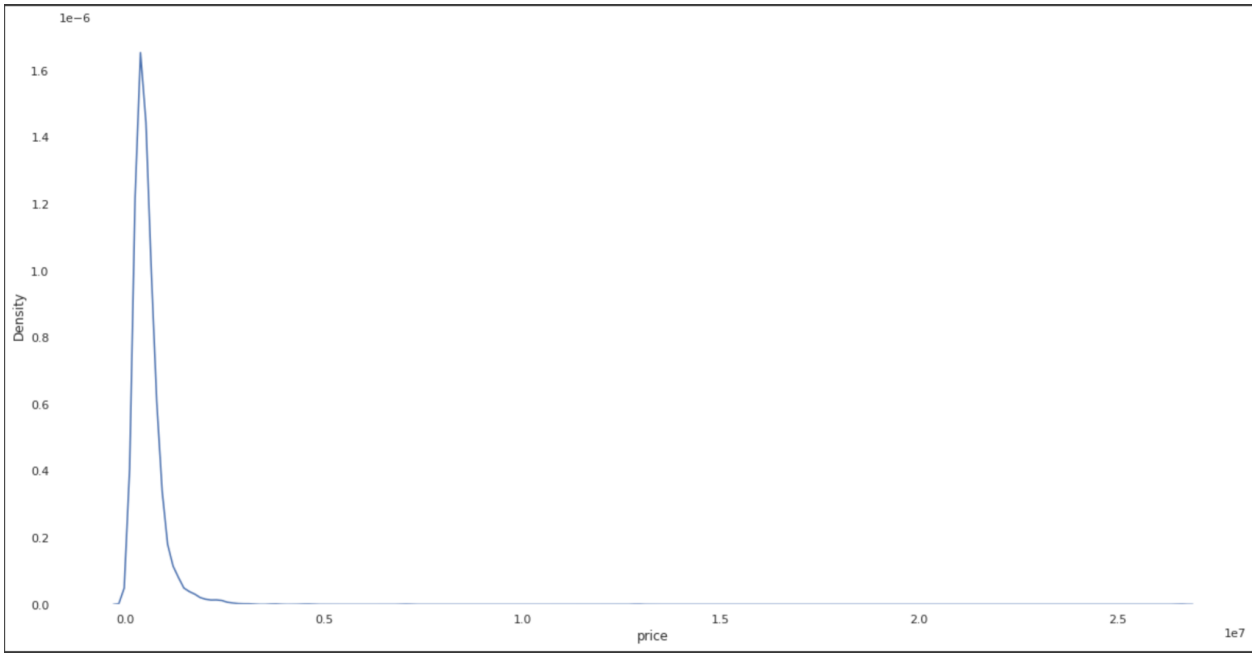


The Price for various cities is also displayed

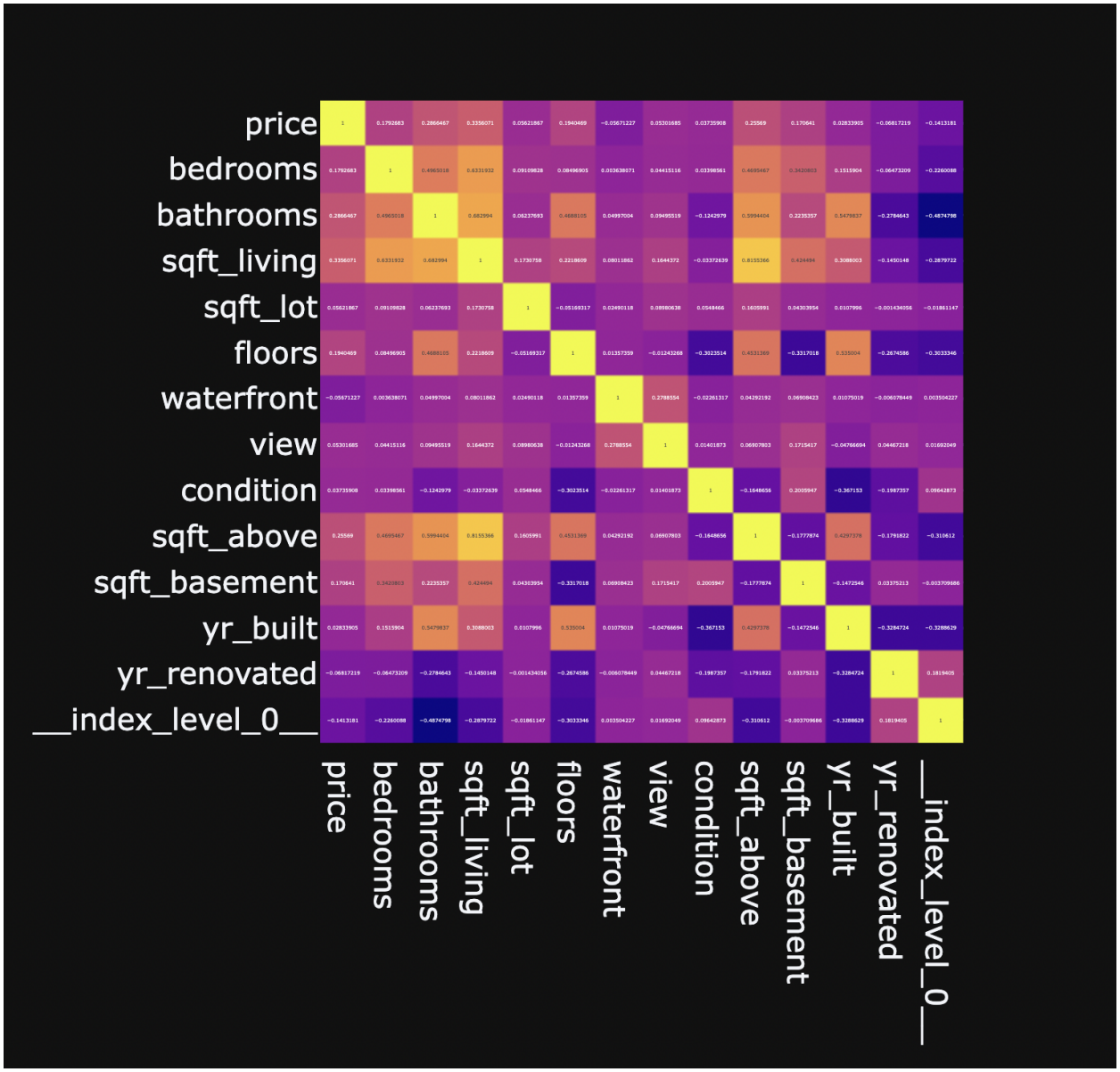


Html file for the live plot can be found here Link

Then drawed a KDE plot on the prices column



The correlation between the numerical columns and price is visualized
Html file for the live plot can be found here Link




```
corr = df_house_price_data.corr(method ='pearson')

f,ax=plt.subplots(figsize=(9,6))
sns.heatmap(corr,annot=True,linewidths=1.5,fmt='.2f',ax=ax)
plt.show()
```

Graph 2:

Drawn some sample plots showcase the relationship between some Numerical datatypes in the dataset with the house price

(Attaching drive link as this alone consists of 10 pdf files)
Images : https://drive.google.com/drive/folders/1XBX2s4q6DZfAJL4fBSTjvvHfJG9zpIgC?usp=sharing

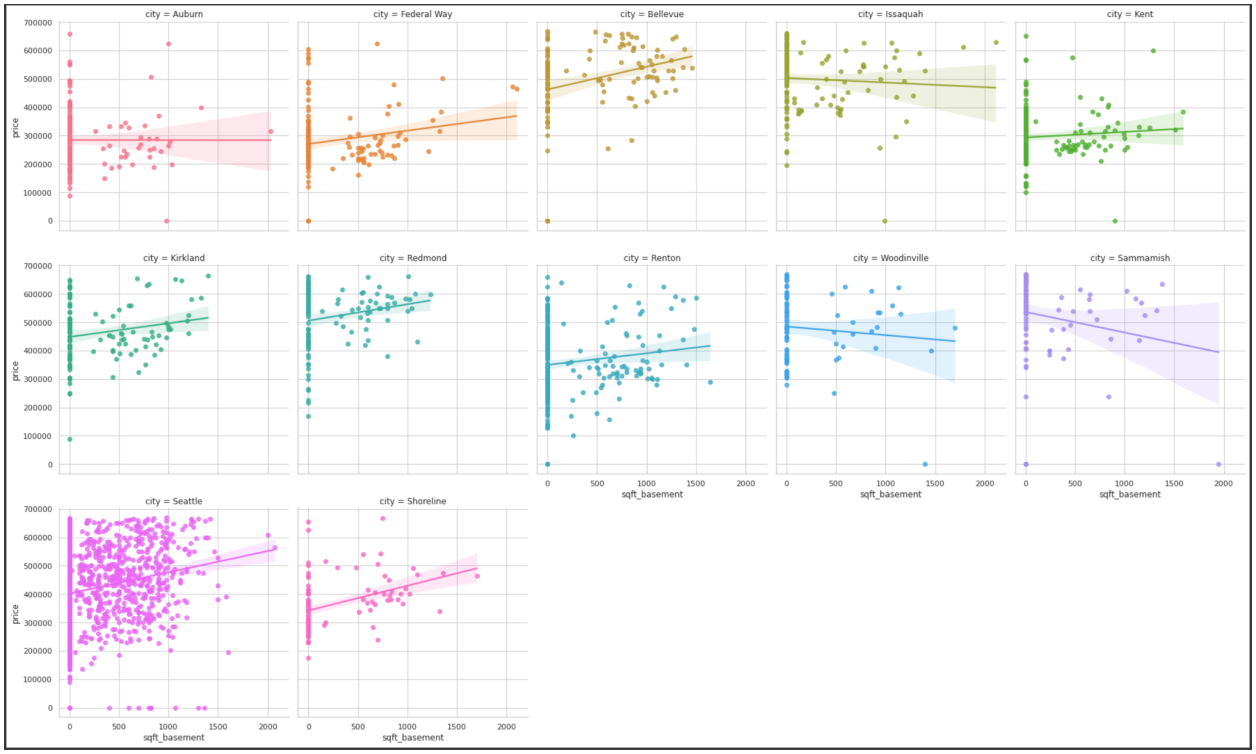
Graph 3:

After this step preprocessed the entire dataset by removing some records where the house price or number of bedrooms or bathrooms or area of the house is not available or is NaN and then reduced the data further by removing the data related to some cities like Yarrow Point as there are not enough records for this city and then further removed the records with prices greater the the 0.75 quantile to minimize the outliers and visualize the data.

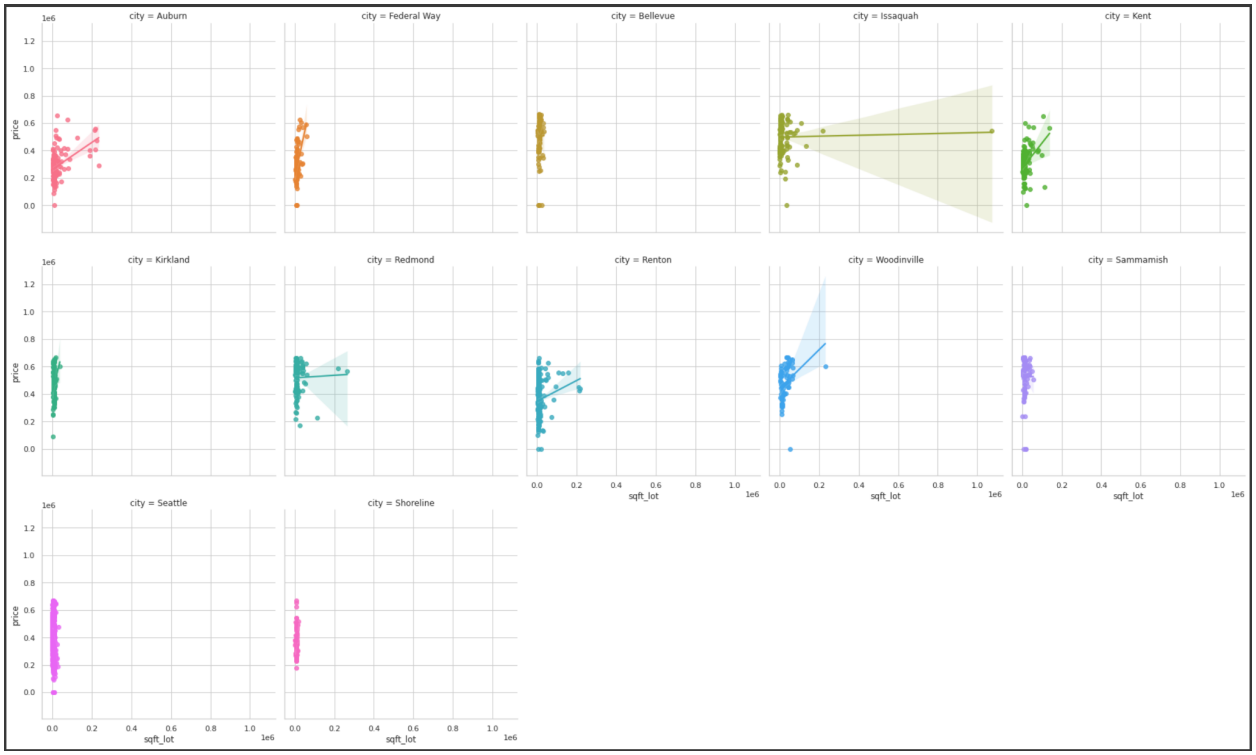
```
quantile = filtered_house_price_df["price"].quantile(0.75)
house_data_final = filtered_house_price_df[(filtered_house_price_df["price"] < quantile)]
filtered_house_price_df.shape
```

Attaching the pdf file links
https://drive.google.com/drive/u/0/folders/11jSLFxpVG_oPtSqDafQ0DKOuRIpwwaGV

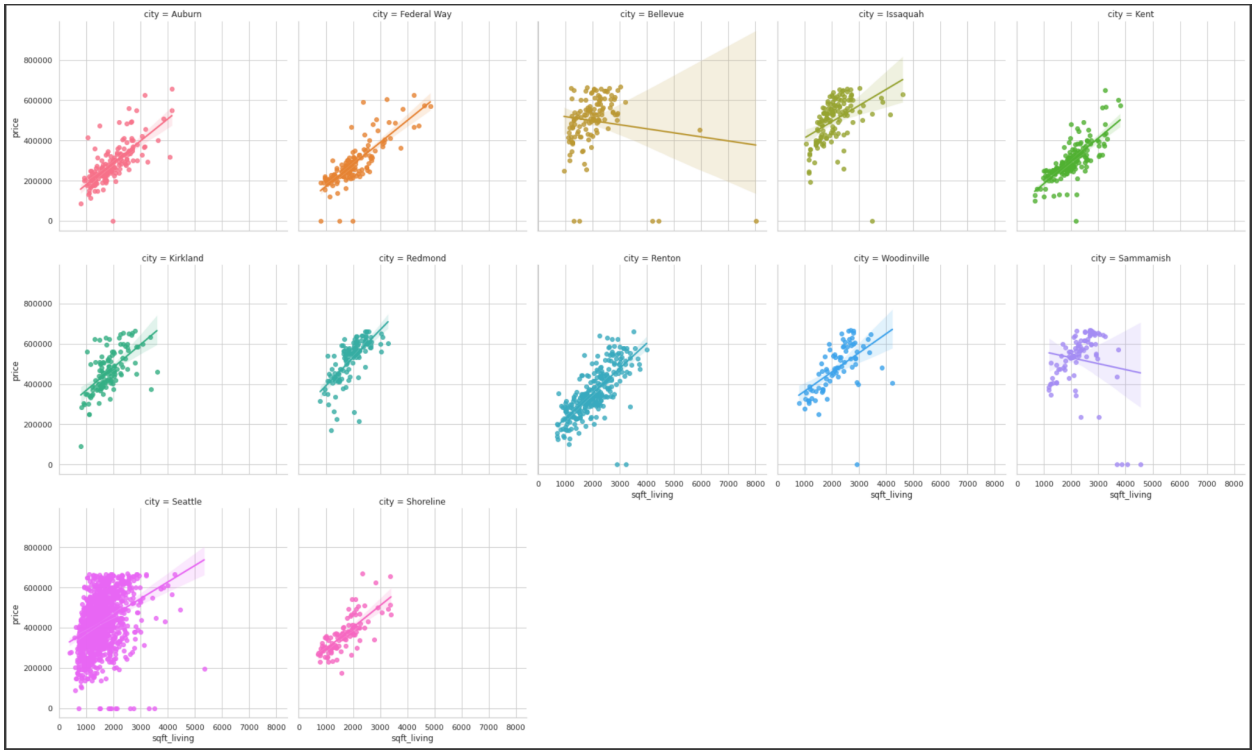
Basement Area vs Prices for all the cities:



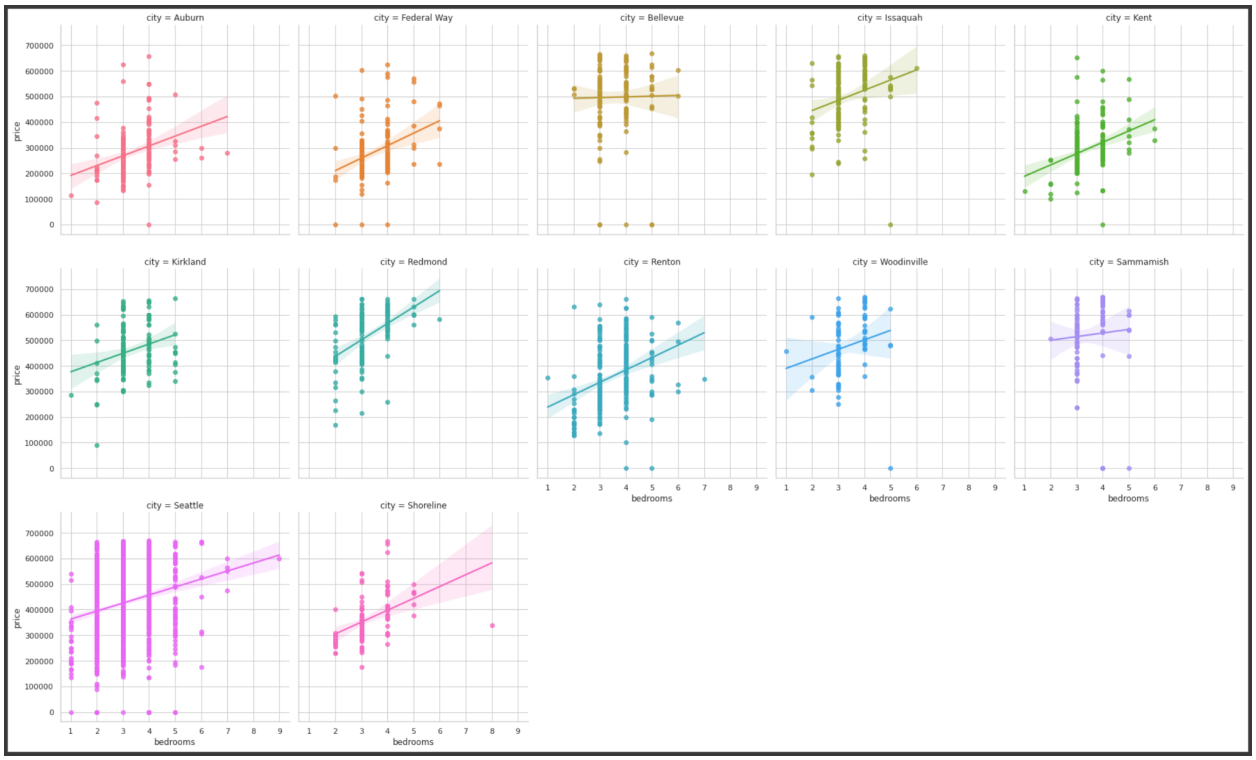
Lot Area vs Prices for all the cities:



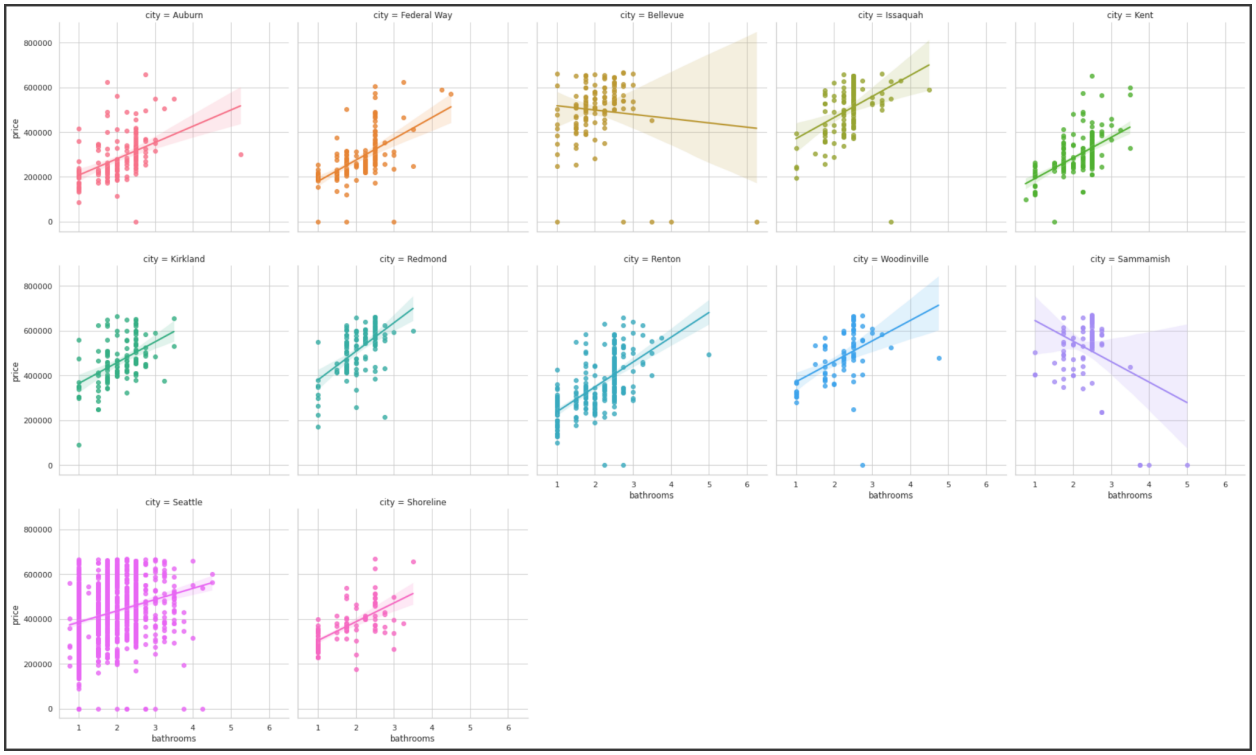
Living Room Area vs Prices for all the cities:



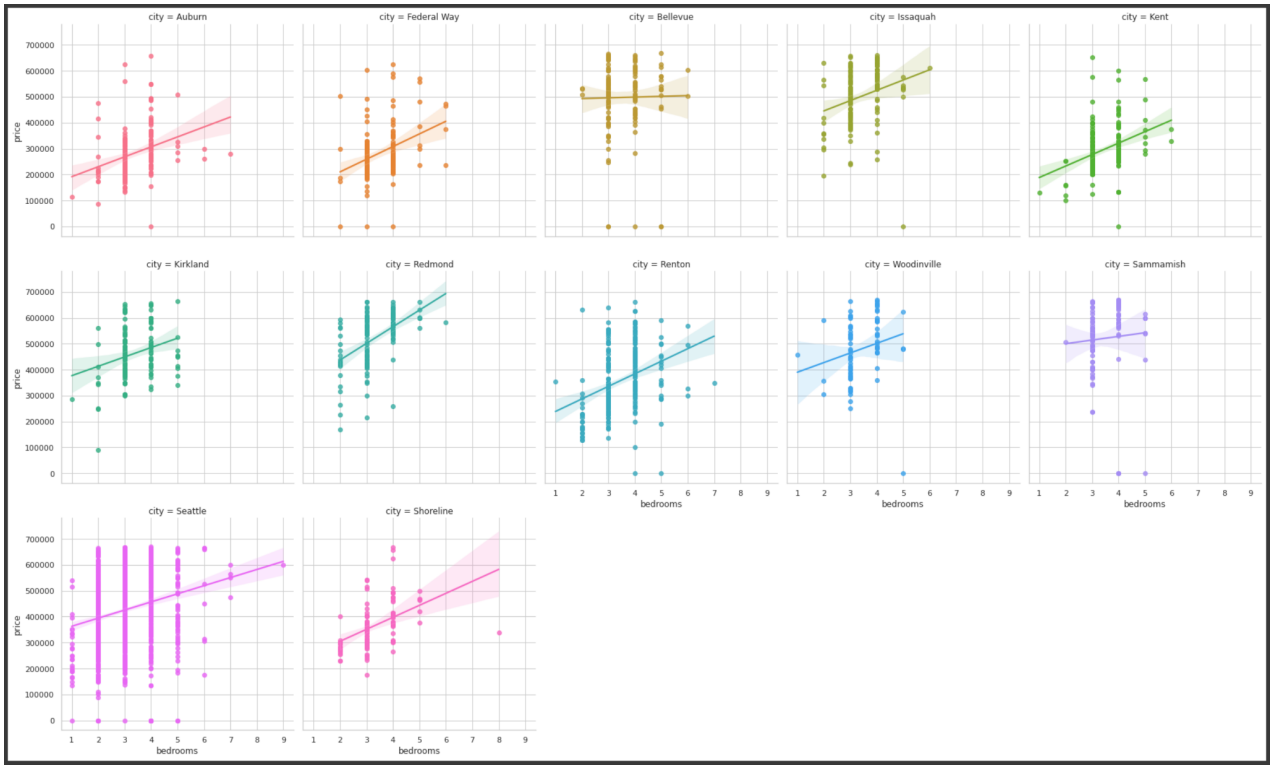
Number of Bathrooms vs Prices for all the cities:



Number of Bathrooms Vs Prices

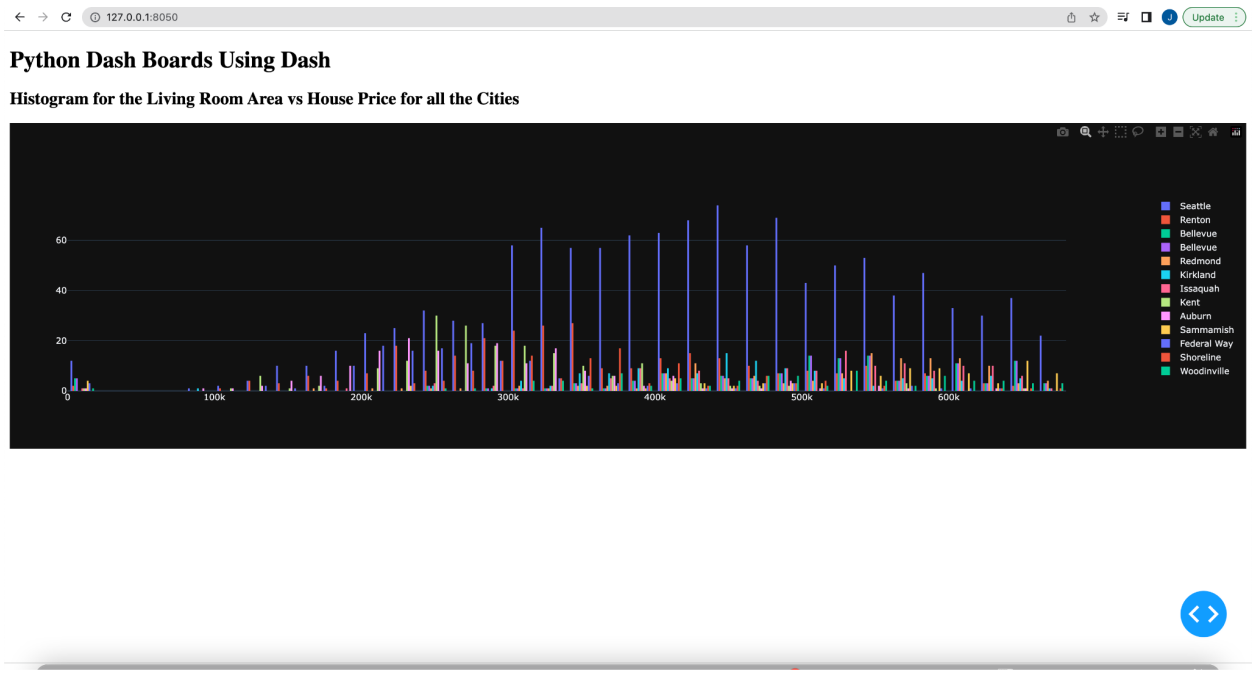


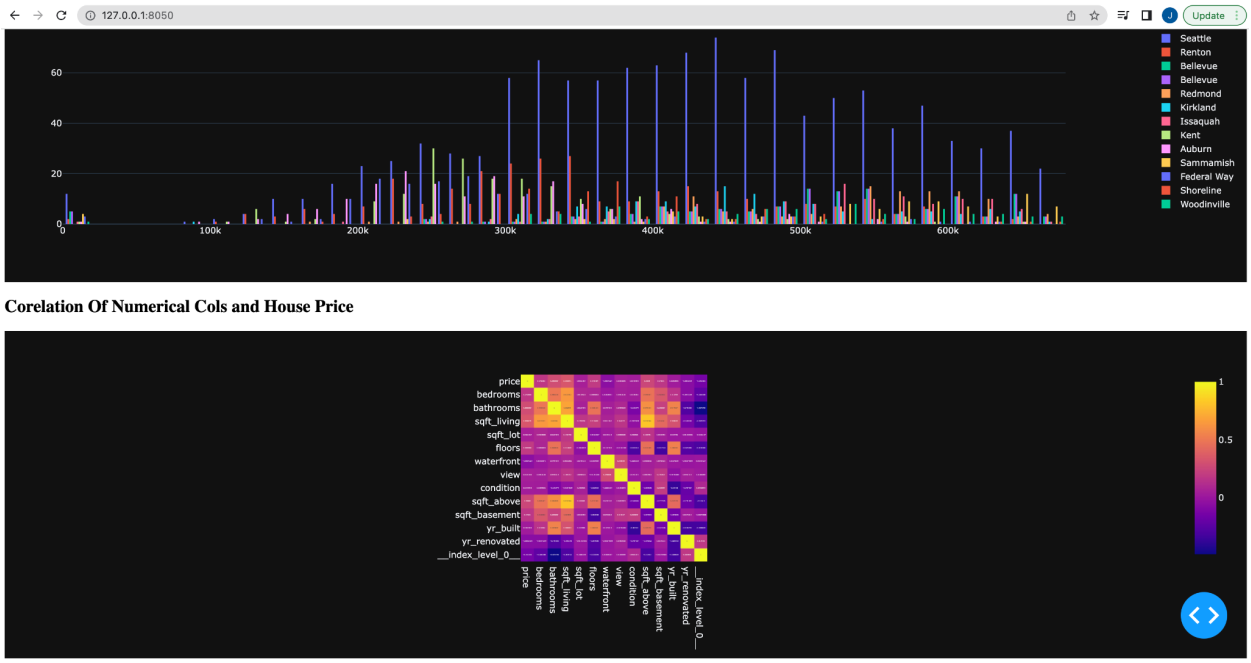
Number of Bedrooms Vs Price Graphs for all the cities



Create A Dashboard using Python Dash:

Use the dm.py file to start the dashboard server and visualize the charts





Upload the final dataset to bigquery:

After performing the simple analysis and filtering the data from the outliers the final data is pushed back to the bigquery dataset in the same project from the google colab notebook
And then tried to retrieve the newly updated final data successfully

```
[ ] for dataset in client.list_datasets():
      print(dataset.dataset_id)
```

```
house_price
```

```
output_dataset_id = 'house_price'

output_table_id = 'preprocessed_data'

replace_or_append_output = 'replace'


project_dataset = (client.project + '.' + output_dataset_id)

# Check to make sure output dataset exists, create it if not
try:
  client.get_dataset(output_dataset_id)
  print("Dataset " + project_dataset + " exists\n")

except:
  print("Dataset " + project_dataset + " doesn't exist, so creating it\n")
  dataset = client.create_dataset(bigquery.Dataset(project_dataset))
```

```
Dataset datamining-364220.house_price exists
```

```
[ ] job_config = bigquery.LoadJobConfig()


    if(replace_or_append_output == 'replace'):
     job_config.write_disposition = bigquery.WriteDisposition.WRITE_TRUNCATE
    else:
     job_config.write_disposition = bigquery.WriteDisposition.WRITE_APPEND

    dataset_ref = client.dataset(output_dataset_id)
    table_ref = dataset_ref.table(output_table_id)


    client.load_table_from_dataframe(
     dataframe = house_data_final,
     destination = table_ref,
     job_config = job_config
     ).result()
```

```
print('Write to bigquery dataset (' + replace_or_append_output + ') to ' + project_dataset + '.' + output_table_id +'\n')
```

```
/usr/local/lib/python3.7/dist-packages/google/cloud/bigquery/_pandas_helpers.py:275: UserWarning: Unable to determine type of column 'street'.
  warnings.warn(u"Unable to determine type of column '{}'.".format(column))
Write to bigquery dataset (replace) to datamining-364220.house_price.preprocessed_data
```

```
sql = """
SELECT
    *
FROM
    `datamining-364220.house_price.preprocessed_data`
"""
house_price_query = client.query(sql)
house_price_df=house_price_query.to_dataframe()
house_price_df.head()
```

1 to 5 of 5 entries Filter

index	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	sqft_above	sqft_basement	yr_built	yr_renovated	street	city	statezip	country	__index_level_0__
0	2014-06-25 00:00:00+00:00	280000.0	7.0	2.5	1940	5458	2.0	0	0	3	1940	0	1994	0	12307 SE 315th Pl	Auburn	WA 98092	USA	1744
1	2014-07-03 00:00:00+00:00	340000.0	8.0	2.75	2790	6695	1.0	0	0	3	1470	1320	1977	2004	17512 Corliss Ave N	Shoreline	WA 98133	USA	2283
2	2014-06-23 00:00:00+00:00	350000.0	7.0	3.0	2800	9569	1.0	0	2	3	1400	1400	1963	2008	2023 Harrington Pl NE	Renton	WA 98056	USA	2381
3	2014-05-15 00:00:00+00:00	475000.0	7.0	3.5	2870	29699	1.0	0	0	3	1520	1350	1961	2004	11738 1st Ave NE	Seattle	WA 98125	USA	2751
4	2014-05-13 00:00:00+00:00	550000.0	7.0	4.0	3440	8100	2.0	0	0	3	3440	0	1970	2014	718 N 95th St	Seattle	WA 98103	USA	2832

Show 25 per page

```
[ ]
```

```
[ ]
```

```
[ ]
```

```
[0]
```

```
[ ]
```