CAPSTONE PROJECT INTERIM REPORT

INTRODUCTION:

In this project, our aim is to predict house prices in certain places of Washington area to give very clear perspective to both sellers and buyers.Prediction of property prices is becoming increasingly important and beneficial. Property prices are a good indicator of both the overall market conditions and the economic health of a country.

STAKEHOLDERS: It is important for real estate markets can categorize house depend on the prices of customer budget to match right customers to required homes while buyers to find home price range that they are looking for.

DATA ACQUISITION:

I acquired the data from Kaggle.(https://www.kaggle.com/shree1992/housedata)

STEPS FOR SOLUTION

1) **Data Wrangling:**
   - Gaining insight about data and taking action according to data itself with Python and Pandas.
   - Checking the data to clean with preprocessing and converting raw data to desired format to be used for future processes.
   - Trying to see any outliers that don't affect data pretty much.

2) **Exploratory Data Analysis(EDA):**
   - Focusing on especially remarkable data that will be detected by Pandas.
   - Deeping into behind to reasons by visualizations.
   - Getting benefit of visualizations to reach the goal using Seaborn and Matplotlib libraries.

3) **Statistical Approach:**
   - Featuring especially things will be searched in previous steps (Data Wrangling and EDA) using Z-T table with OLS chart
   - Working on p-values of estimations in main and support branch of this capstone.

DATA WRANGLING:

DATA: House data has 4600 rows with 18 different columns.Set_index,groupby, concatenate, pivot_table are some of pandas techniques I applied in Data Wrangling parts.

LOADING DATA: Lets load the data and try to see general view of it

	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	sqft_above
0	2014-05-02 00:00:00	313000.0	3.0	1.50	1340	7912	1.5	0	0	3	1340
1	2014-05-02 00:00:00	2384000.0	5.0	2.50	3650	9050	2.0	0	4	5	3370
2	2014-05-02 00:00:00	342000.0	3.0	2.00	1930	11947	1.0	0	0	4	1930
3	2014-05-02 00:00:00	420000.0	3.0	2.25	2000	8030	1.0	0	0	4	1000
4	2014-05-02 00:00:00	550000.0	4.0	2.50	1940	10500	1.0	0	0	4	1140

CAPSTONE PROJECT INTERIM REPORT

OUTLIER:

Are there homes with $0 prices?

There are 49 homes with $0 that we can drop these columns that doesn't affect our dataset with dropna method.Our data has 4600-49=4551 features.

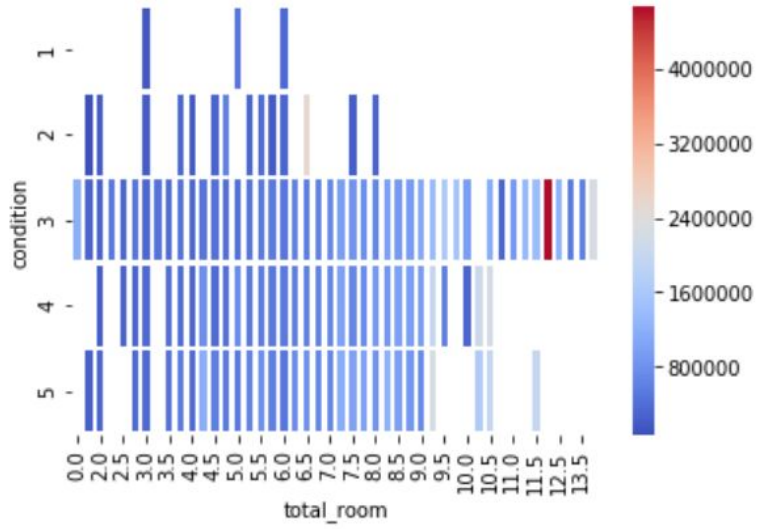
```
len(house[house.price==0])
```

```
49
```

In this graph below, we can see the triple relationships among price-#total room-condition.There is a home with dark orange color which condition is 3 and its price is pretty bigger than average.This home price needs to be reviewed in detail. We see the same issue for the home with condition is 2.It is also way above the average and required what the reason behind is.

```
df = h.pivot_table(index='condition',columns='total_room',values='price')
sns.heatmap(df,cmap='coolwarm',linewidths=1.5)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x22f03764dd8>
```



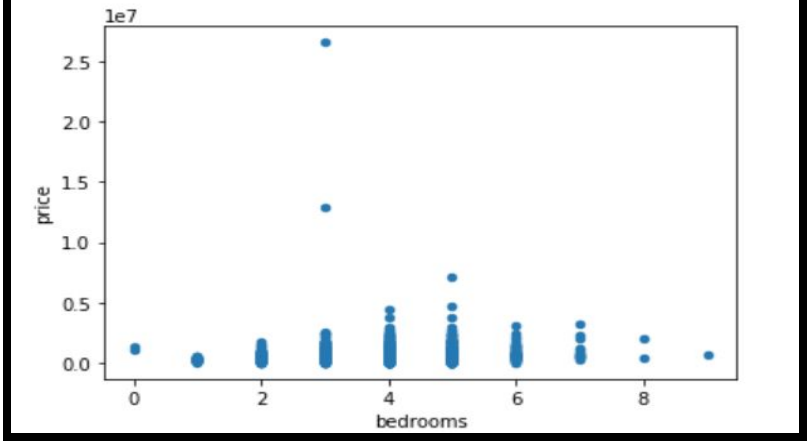
EXPLORATORY DATA ANALYSIS (EDA):

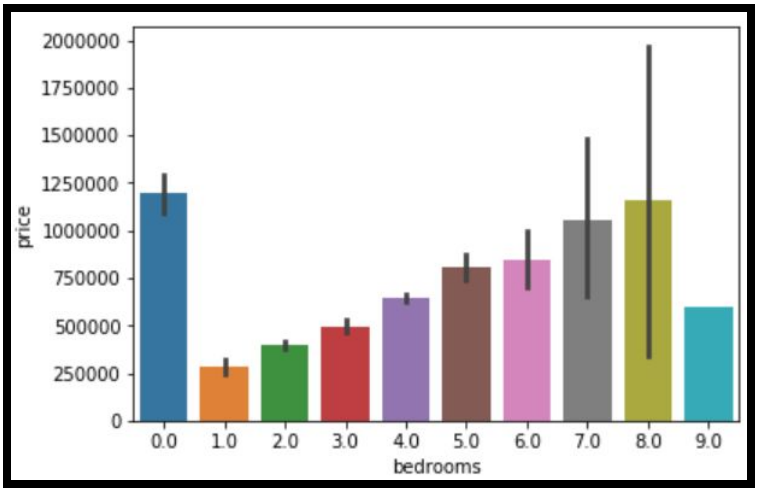
GRAPHS AND TYPES:(Histogram,Scatter,Bar,Heatmap)

CAPSTONE PROJECT INTERIM REPORT

We can easily have a chance to see home prices with outliers scattering depend on bedroom numbers.

```
house.plot(kind='scatter',x='bedrooms',y='price')
plt.show()
```

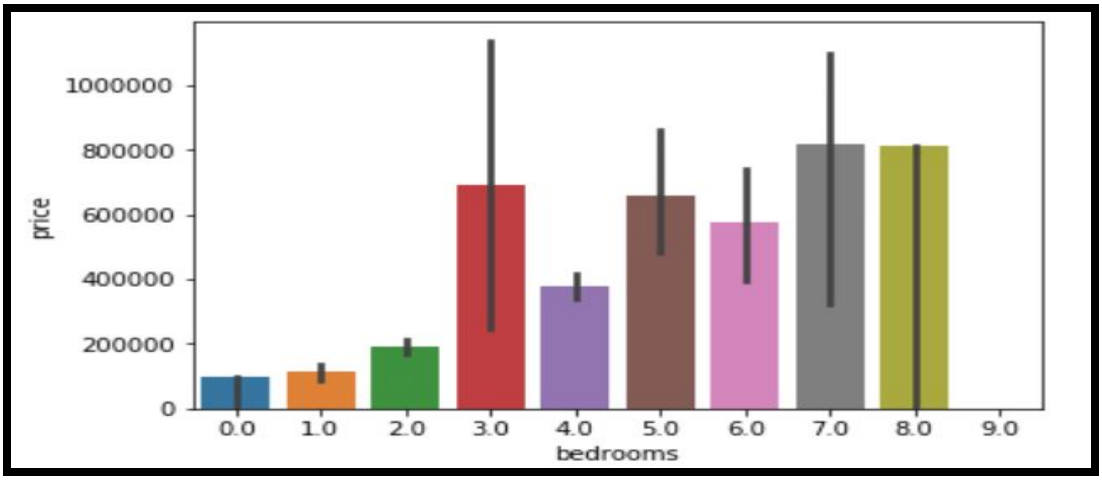




If we desire to see the same features on houses above standards.

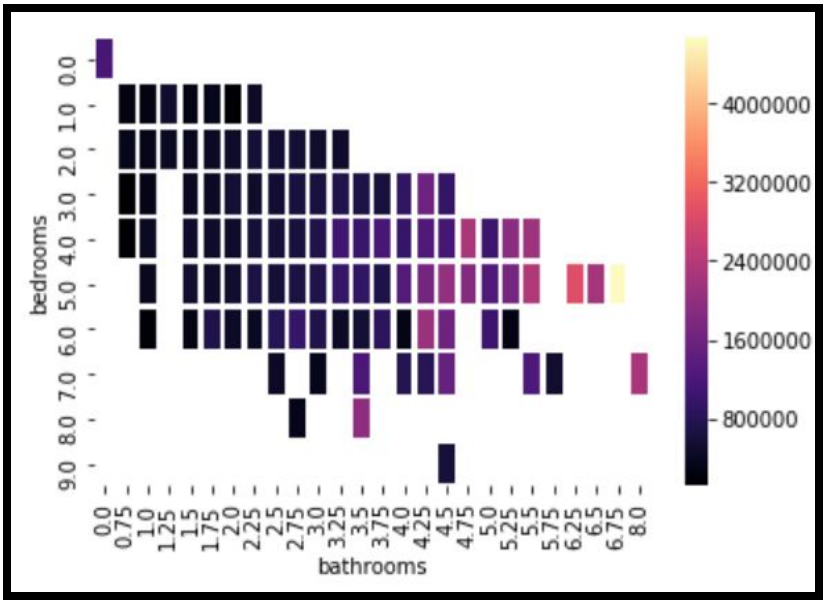
```
h=new_house
h[h.price >= new_house['price'].mean()]
h
sns.barplot(x='bedrooms',y='price',data=h,estimator = np.std)
```

CAPSTONE PROJECT INTERIM REPORT



Let's use heatmap and pivot_table to get an idea of effecting price both bedroom and bathroom numbers.

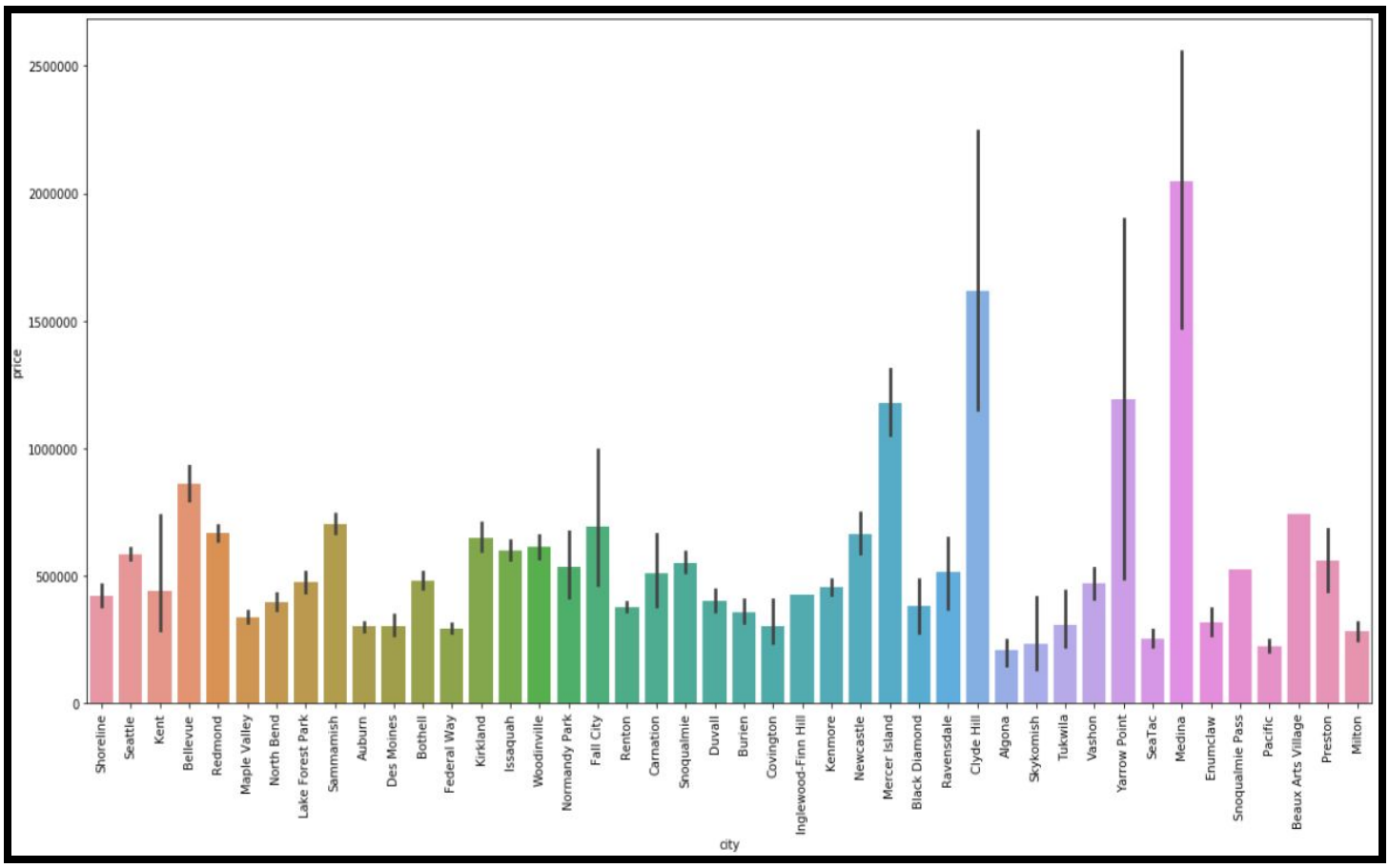
```
df = h.pivot_table(index='bedrooms',columns='bathrooms',values='price')
sns.heatmap(df,cmap='magma',linewidths=1.5)
```



CAPSTONE PROJECT INTERIM REPORT

EDA RESULTS:

Some homes with 0 rooms might be a studio or office in the center of the city. There are some distinct zip codes that have higher prices, as well as the opposite situation exists may depend on safety,crime rates or closing to important places. Good conditions,perfect views also contribute positive side to the price. Number of bedroom and bathroom will huge impact on prices. People in interestingly consider bathroom number. Prices with under marketing should be either foreclosure homes,short sales,fixer uppers, homes in bad locations,homes priced too low by mistakes.



CAPSTONE PROJECT INTERIM REPORT

STATISTICAL APPROACH:

Statistical approach is a powerful way at catching significant data and helping machine learning algorithms to build an effective model.

In this chapter, I will continue my capstone project about House predictions approaching statistically way.Here are my three different observations:

1) Comparing average price of homes with 2 and 3 bedrooms.
2) Determining average price in Issaquah greater than average price in Kirkland or not
3) Probability of a home hired by a family just moves into this area.

Inquiry:

A mortgage company wants to test if there is a significant difference between the mean price of homes with 3 bedrooms and the mean price of homes with 2 bedrooms in given dataset.

(significance level=0.05)

Let M2 and M3 be mean of homes that have 3 bedrooms and 2 bedrooms respectively.
H0: M2 = M3 (M2 - M3 = 0) and H1: M2 ≠ M3 (M2 - M3 ≠ 0)

EXAMINING HOUSES WITH 3 AND 2 BEDROOMS:

```
len(house[house['bedrooms']==3])
```

```
2032
```

```
h3 = house[house['bedrooms']==3]
house_3 = h3['price']
```

```
round(np.mean(house_3),2)
```

```
488613.02
```

```
round(np.std(house_3),2)
```

```
690338.86
```

```
len(house[house['bedrooms']==2])
```

```
566
```

```
h2 = house[house['bedrooms']==2]
house_2 = h2['price']
```

```
round(np.mean(house_2),2)
```

```
391621.92
```

```
round(np.std(house_2),2)
```

```
194947.17
```

let's take a sample from both house type that contains 50 houses from M2 and M3 respectively.(50<(566/10)),(50<(2032/10)) For the sake of 10% rule being independent, 50 is a plausible value which is less than 10% of the number of houses that have 2 and 3 bedrooms.

CAPSTONE PROJECT INTERIM REPORT

In this frequentis statistic,z value is way beyond z_critical value so we have to reject the null hypothesis(M3-M2=0) which we accept the alternative hypothesis. It indicates that there is a significant difference between mean price of homes with 2 bedrooms and mean price of homes with 3 bedrooms.

```
# I am finding mean and standard deviation of the sample respectively for houses with 2 bedrooms.
```

```
seed(35)
sample_2 = np.random.choice(house_2,50)
sample_mean_2 = np.mean(sample_2)
round(sample_mean_2,2)
```

```
425348.56
```

```
seed(35)
sample_std_2 = np.std(house_2)/(50**0.5)
round(sample_std_2,2)
```

```
27569.69
```

```
#I will find mean and standard deviation of the sample belongs to houses with 3 bedrooms.
# For the sake of 10% independent rule, I will choose 50 houses.(50<(2032/10))
```

```
seed(35)
sample_3 = np.random.choice(house_3,50)
sample_mean_3 = np.mean(sample_3)
round(sample_mean_3,2)
```

```
539394.89
```

```
seed(35)
sample_std_3 = np.std(house_3)/(50**0.5)
round(sample_std_3,2)
```

```
97628.66
```

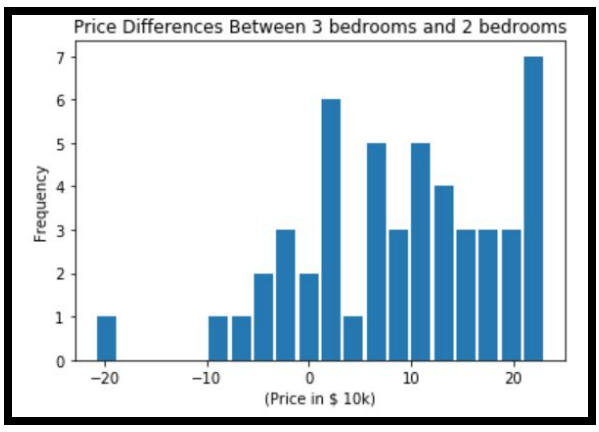
Now let`s apply z test

```
z = (sample_mean_3-sample_mean_2)/np.sqrt(np.sum((sample_std_2**2/50**0.5)+(sample_std_3**2/50**0
.5)))
round(z,2)
```

```
2.99
```

```
#significance level = 0.05
z_critical = norm.ppf(0.975)
round(z_critical,2)
```

```
1.96
```



CAPSTONE PROJECT INTERIM REPORT

2) Is average price in Issaquah greater than average price in Kirkland?

We will inquiry house prices in two same size cities, Kirkland and Issaquah(both city has 187 homes) if the average price of homes in Issaquah greater than average price of homes in Kirkland. If I investigate having normal distribution conditions:

1) 187 homes were randomly chosen homes from both cities separately.

2) 187 homes for Kirkland is below 10% of total homes.(Independent) (Apparently number of homes in Kirkland is near 22000) same thing valid for Issaquah.187 homes is quite under the total number of homes(10000) in it. 3) Our sample is reasonably large 187>30(for both samples)

```
Kirkland = house[house.city == 'Kirkland']
round(np.mean(Kirkland['price']),2)
```

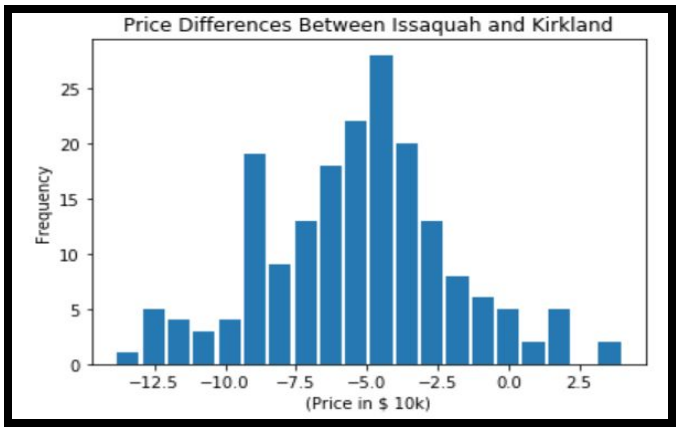
```
round(np.std(Kirkland['price']),2)
```

```
Issaquah = house[house.city == 'Issaquah']
round(np.mean(Issaquah['price']),2)
```

```
round(np.std(Issaquah['price']),2)
```

```
#Let`s start finding with z in order to see whether it falls into area within 95% interval or not.
z = (596163.75-651583.59)/np.sqrt(np.sum((245008.93**2/187)+(368857.02**2/187)))
round(z,2)
```

Now it is time to check falling this z value into within 95% confidence interval or not. For this reason, we need to figure our p value out.(significance level = 0.05) P value of -1.71 on z table is 0.0436 which is smaller than the significance level(0.05). We will reject the Null Hypothesis in favor of alternative hypothesis.It means average price of Homes in Issaquah is not expensive than Kirkland.



In other words, home prices in Kirkland is more expensive than home prices in Issaquah.

As you can see on the graph,major data gathered below 0 which mean of home prices in Issaquah is lower than mean of home prices in Kirkland.

CAPSTONE PROJECT INTERIM REPORT

3) BAYES THEOREM

Wederson Family moves from California to Washington. They search for a home that has 2 bathrooms and 3 or 4 bedrooms in new area.

What is the probability of they rent a home with condition is not lower than 4 (with Ms. Wederson`s demand) of given options above?(Solve by using Bayes Theorem)

X: probability of homes with 3 or 4 bedrooms and 2 bathrooms Y: probability of homes with not 3 or 4 bedrooms and not 2 bathrooms A: probability of condition greater than or equal to 4 B: probability of condition less than 4 and Mr. Wederson is searching among the homes that have 3 or 4 bedrooms and 2 bathrooms.(X) Of those homes,with his wife `s request, he will try to find out a home condition is greater than or equal to 4.(A)

Statistically,he is seeking P(A/X)

P(A|X)=(P(X|A)*P(A))/P(X) (BAYES T.)

Question: How many homes with 3 or 4 bedrooms and 2 bathrooms are there?

```
X = house[((house.bedrooms==3)|(house.bedrooms==4))&(house.bathrooms==2)]

print('P(X):{}{}{}'.format(len(X),'/',4600))
```

```
P(X):355/4600
```

Question: How many home conditions does have greater than or equal to 4?

```
A = house[house.condition>=4]

print('P(A):{}{}{}'.format(len(A ),'/',4600))
```

```
P(A):1687/4600
```

```
#Let`s find out of up_4 how many homes are there with 3 or 4 bedrooms and 2 bathrooms.
#P(X|A)=?
X_A = A[((A.bedrooms==3)|(A.bedrooms==4))&(A.bathrooms==2)]
print('P(X|A):{}{}{}'.format(len(X_A),'/',1687))
```

```
P(X|A):177/1687
```

Now, I am plug in values into Bayes formula

P(A|X)=(P(X|A)*P(A))/P(X) (BAYES T.)

```
(((177/1687)*(1687/4600))/(355/4600))*100
```

```
49.859154929577464
```

CAPSTONE PROJECT INTERIM REPORT

RESULT:

Probability of Wederson Family choosing homes they wish is 49.9%.

It seems they are lucky 1 out of 2 homes is the home they are looking for and they can easily access or find those homes by filtering method in any home selling websites.