GROUP : 9

- Raghav Goel
- Leroy Rebello
- Rachita Kadam
- Aishwarya Gabhane
- Mihir Patel

REAL ESTATE ANALYSIS 

Introduction and Web Scraping

- Redfin – Real Estate Marketplace
- Link: https://www.redfin.com/
- Our Focus on Northern California Listed Properties
- Type of Data : Float, Int, String(Object)

Preliminary Ideas of Analysis*

- Exploratory data analysis
- City/County wise Avg price differences
- How number of beds/baths vary city/county wise
- Top 5 cities with maximum property rates
- Correlation between attributes
- Observations & Insights
- Prediction Analysis

*Subject to Change

Web scraping Process

1. List of County URLS
2. Extracted Webpages of each county
3. Dictionary of Properties
   Key : County
   Value : List of properties
4. Fetching property information
5. Converted data to Data Frame and cleaned the data.
6. Exported to CSV and ready for analysis!

Web Scraping Process in Brief

1 List of County URLs

- Initial list of counties with first webpage of each county

2 Extracted all webpages of each county

- Used BeautifulSoup to find the number of webpages and add all to the list.

3 Dictionary of properties

- Key - County
- Value - List of properties
- a href tags extracted and used regex to filter for only properties and removed duplicate links.

4 Property information

- Lists created for attributes - Street, City, etc
- Created a BS object for property link and extracted information using different relevant tags like 'div', 'span', etc.
- Used try & except blocks to handle errors.

5 Converted data to DataFrame and cleaning the data

- Used pandas DataFrame
- Identified the null values, outliers, noisy data and cleaned the data

6 Exported to CSV and ready for analysis!

- Data is ready for use!

Challenges Faced

- Blocked by Redfin (We were considered as robots and blocked by the website to web scrape)
- Could Extract the first page of county listing. Used span tag and page text class to get all the available listings .
- Duplicate Links for the same property.
- Empty plot (Land) threw error while extracting bed, bath and sqft.
- Same tag and class name being used for different attributes. Became difficult to extract data individually.

DataSet Description

Data Extracted Consists of
8238 rows and 15 columns

Street	Location of the property
City	City of the property
State	State of the property
Zip Code	Zip Code of the property
Price	Listed price/selling price/market value for the property
Bed	Number of rooms in the property
Bath	Number of baths in the property
Sqft	Overall area in square feet
Walk score	Walk Score measures the walkability of any address
Bike score	Bike Score measures whether an area is good for biking
Property Type	Type of Property
Year Built	Year of Built
Status	Current Status - Active, Sold, New, Coming Soon
Acre	Size of the empty lands (No house built)
County	Area in which the property is located in

Data Cleaning Process

Final Data Cleaned Consists of
4661 rows and 14 columns

- Check for Null values and remove
- Check for Duplicate values and remove
- Remove Special Characters
- Normalization
- Changed Data types to relevant ones
- Detect and remove Outliers
- Check for possible null values again and remove

Assumptions

- Only Dealing with Properties that have houses built
- Focus on Northern California

Exploratory Data Analysis

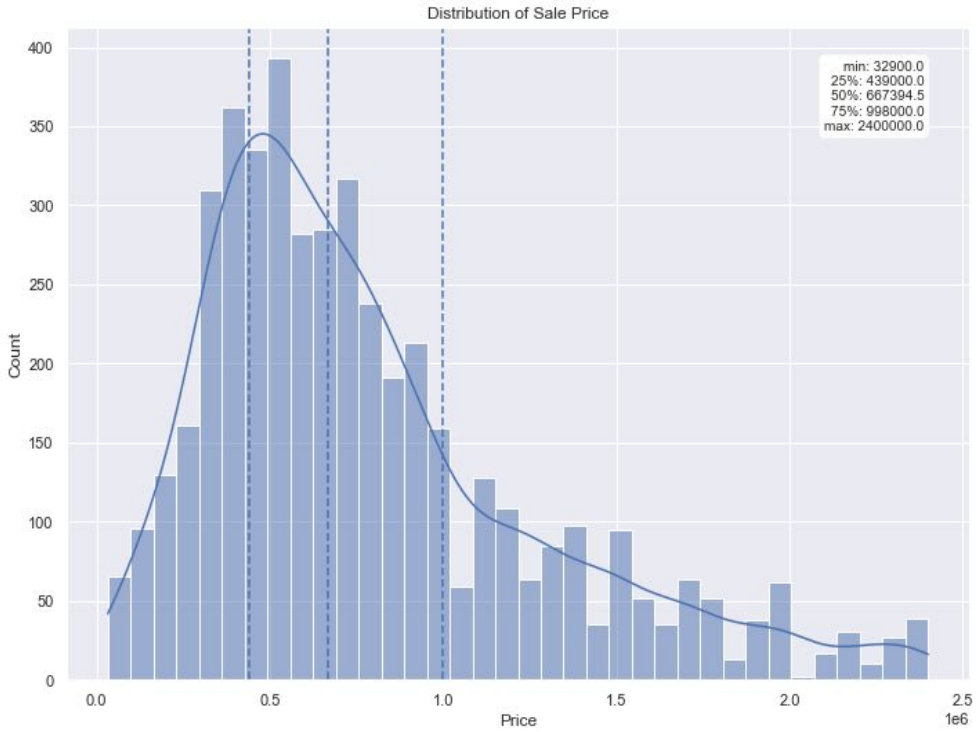
- Performed Exploratory Data Analysis to get a bird eye view of data. It showcases how the data is distributed
- Gives a checkpoint to start other Analysis.

In [5]:
```
df.describe()
# Performed Exploratory Data Analyis to get a bird evye of data. It Showcase ehow the data is distrbuted
# Gives a checkpoint to start other Analaysis.
```

Out[5]:

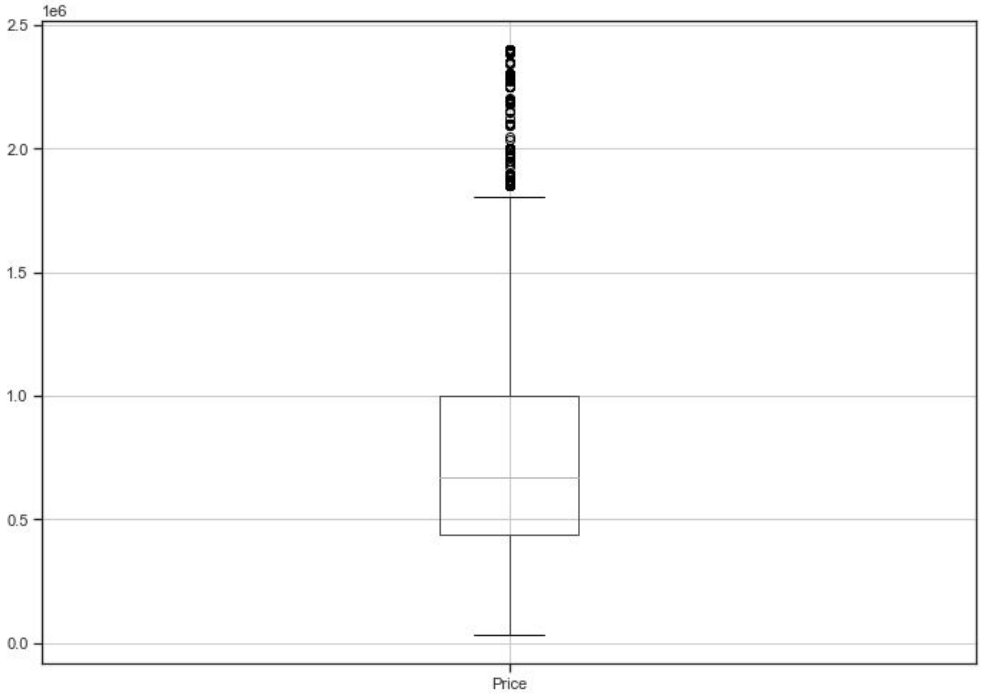
	Price	WalkScore	Bike_Score	Beds	Bath	Sqft	Year_Built
count	4.652000e+03	4652.000000	4652.000000	4652.000000	4652.000000	4652.000000	4652.000000
mean	7.910861e+05	34.526010	39.610275	3.021281	2.171700	1763.235813	1978.531599
std	4.942269e+05	31.142938	26.251453	1.032361	0.758974	744.485024	28.851231
min	3.290000e+04	0.000000	0.000000	0.000000	0.500000	275.000000	1850.000000
25%	4.390000e+05	4.000000	19.000000	2.000000	2.000000	1210.000000	1961.000000
50%	6.673945e+05	28.000000	39.000000	3.000000	2.000000	1636.000000	1981.000000
75%	9.980000e+05	61.000000	59.000000	4.000000	2.500000	2178.000000	2001.000000
max	2.400000e+06	100.000000	100.000000	7.000000	4.500000	4212.000000	2023.000000

Distribution of Price



- Plotted the prices of all the counties to see how well the data is distributed.
- We can conclude that price is positively skewed and mean > mode.
- The positive skewness tells us that there are outliers on the higher range of price of house listing.

Outliers - Positively Skewed

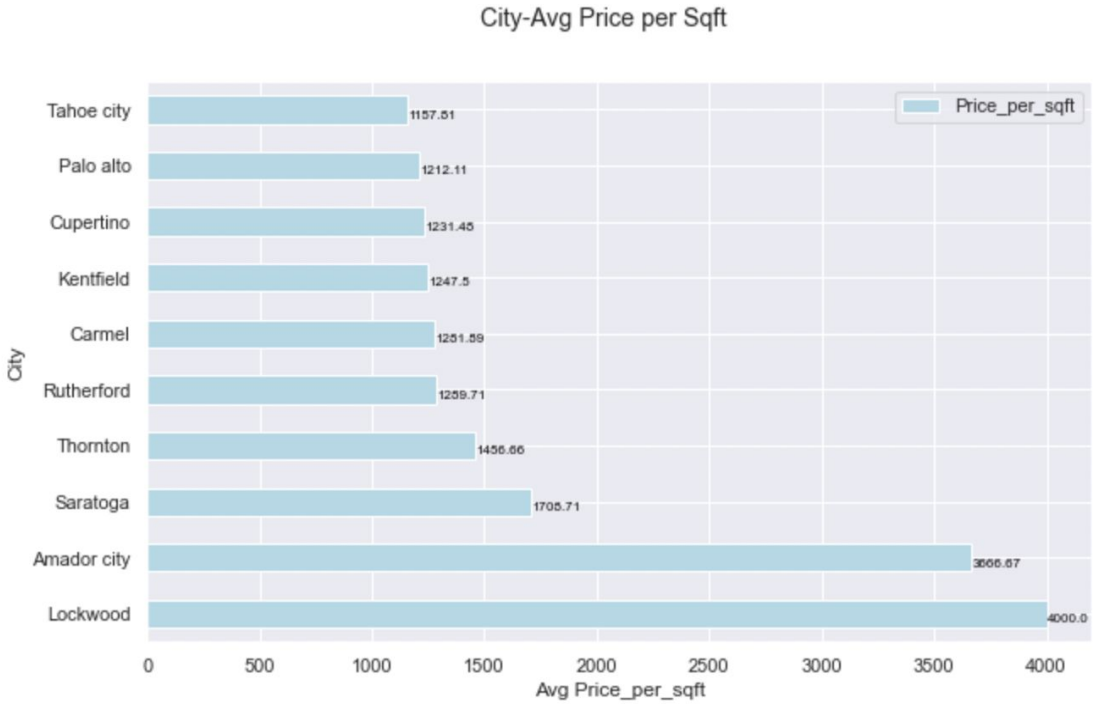


- As the price distribution showed there might be some outliers on the higher hand of the price range.
- To confirm our analysis we plotted Box Plot on price listing. We can safely conclude that there are outliers on the higher price range.

County Wise Walk Score and Bike Score

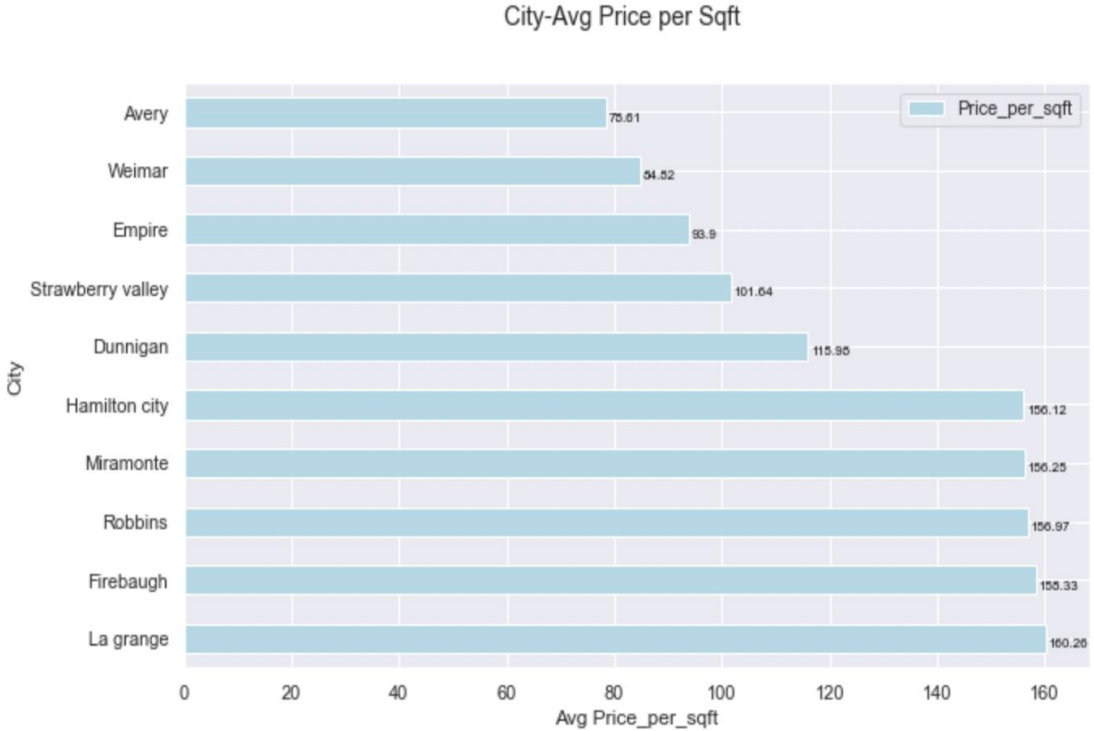
- It shows a bird's eye view of which counties have a higher walk and bike score.
- This will be investigated further in our analysis.

Top 10 cities in CA as per the Average price per square feet



- Lockwood is the most expensive city in california as per price per sq ft.
- Top 10 cities lies between ~$1150-$4000, shows a vast range in dollar amount.
- This analysis also shows that all other cities are comparatively affordable as all are below $1157 per square feet.

Bottom 10 cities in CA as per the Average price per square feet



- Avery is the least expensive city in California as per price per square feet.
- Bottom 10 cities lies between $75 -$160, making it quite affordable compared to other cities.
- Hamilton City, Miramonte and robbins are almost same in their average price per square feet costing.
- Another interesting fact after the analysis implies that 85% of the cities lie between $160 -$1250

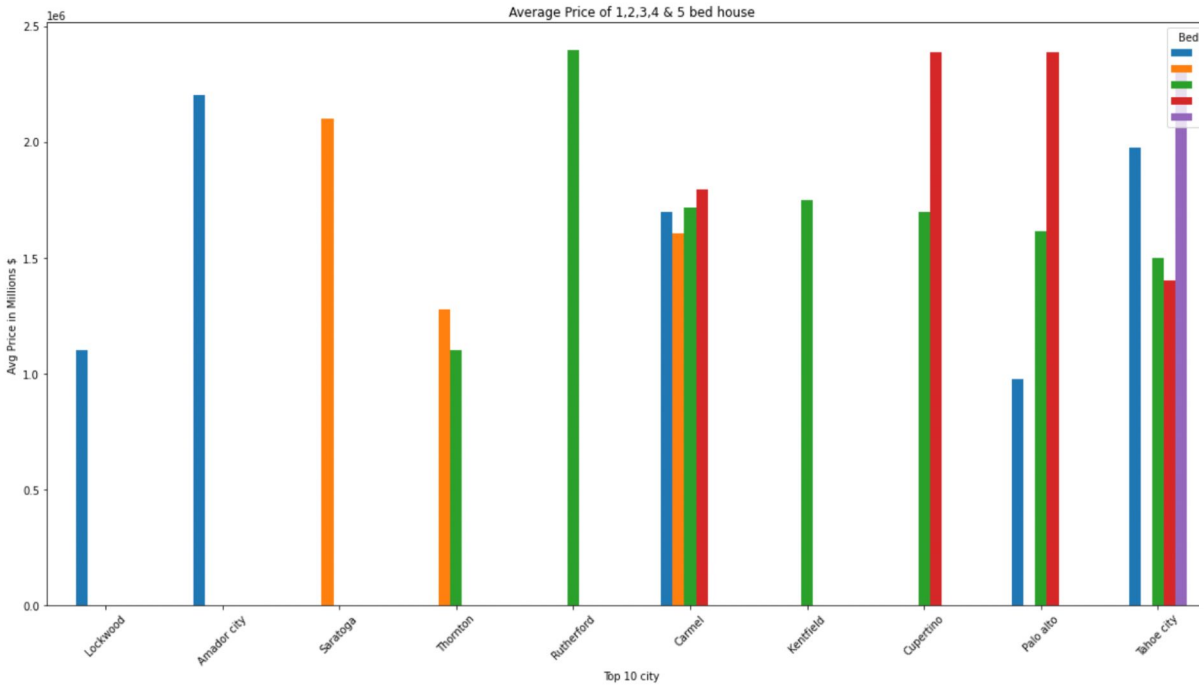
Total Properties as per no of beds

Beds	0.0	1.0	2.0	3.0	4.0	5.0	6.0	7.0
County								
Alameda-County	1	19	93	103	72	14	9	1
Alpine-County	0	8	4	3	2	0	0	0
Amador-County	0	3	22	46	8	1	1	0
Butte-County	0	6	44	84	28	8	1	0
Calaveras-County	2	4	18	50	11	3	0	0
Colusa-County	0	0	5	6	3	0	1	0
El-Dorado-County	0	7	36	121	69	19	6	1
Fresno-County	0	5	38	135	65	21	1	1
Glenn-County	0	0	9	16	10	2	1	0
Lake-County	0	5	57	74	18	3	2	0
Madera-County	0	4	19	78	55	15	3	1
Marin-County	1	21	30	41	11	3	0	0
Mariposa-County	0	0	17	25	8	0	2	0
Monterey-County	0	7	38	66	33	5	3	0
Napa-County	1	2	30	41	17	4	1	1
Nevada-County	0	9	34	110	28	3	0	1
Placer-County	3	5	34	95	84	21	4	0
San-Benito-County	0	0	6	27	17	1	1	0
San-Francisco-County	9	48	98	57	20	6	1	1
San-Joaquin-County	0	7	74	138	73	27	5	1
San-Mateo-County	3	26	74	81	28	6	2	1
Santa-Clara-County	1	13	61	116	61	17	2	0
Santa-Cruz-County	2	9	55	45	21	5	0	0
Solano-County	0	8	49	88	87	19	2	0
Sonoma-County	1	10	63	88	44	8	2	0
Stanislaus-County	0	5	38	136	67	22	1	1
Sutter-County	1	1	5	31	17	4	2	0
Yolo-County	0	3	14	30	25	8	1	1
Yuba-County	0	4	13	46	31	9	1	0

3.0	1977
2.0	1078
4.0	1013
5.0	254
1.0	239
6.0	55
0.0	25
7.0	11

- Total properties as per no of beds across all county
- Further splitting total house count as per number of beds.

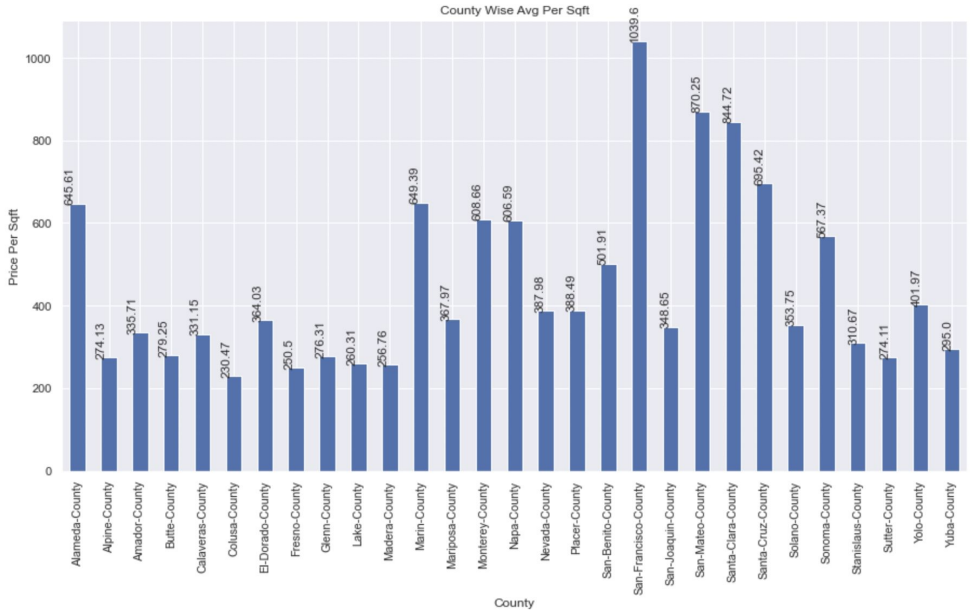
Average prices of house per bed



Beds	1	2	3	4	5
City					
Lockwood	1100000	0	0	0	0
Amador city	2200000	0	0	0	0
Saratoga	0	2100000	0	0	0
Thornton	0	1274950	1100000	0	0
Rutherford	0	0	2395000	0	0
Carmel	1700000	1606000	1714500	1795000	0
Kentfield	0	0	1749000	0	0
Cupertino	0	0	1700000	2388000	0
Palo alto	978000	0	1614294	2388000	0
Tahoe city	1975000	0	1500000	1400000	2300000

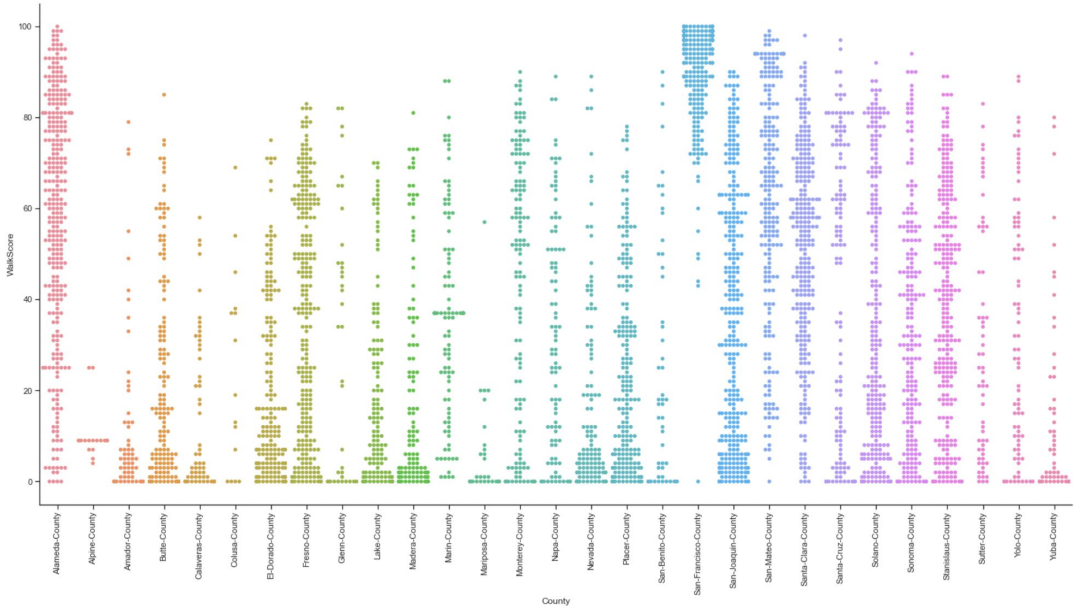
As number of bed increases in a property, the average property Price also increases.

Bar plot of county and average price per sqft



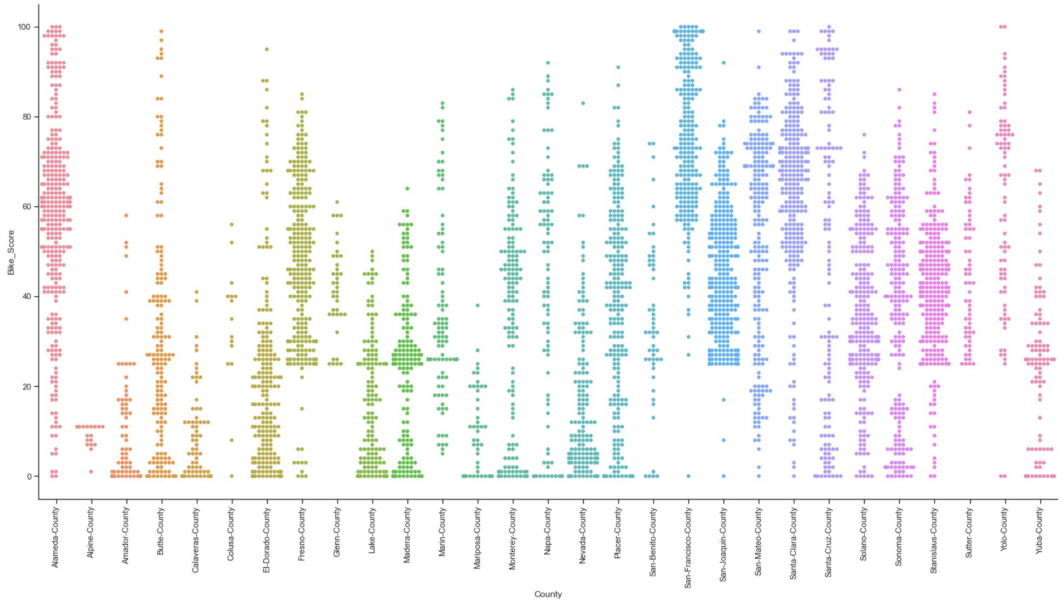
- San Francisco, Santa Clara and San Mateo are three top most expensive counties in California as per Price Per Sqft where San Francisco being the highest.
- Majority of the counties have Price/Sqft below $400.

Catplot of Walk Score County Wise



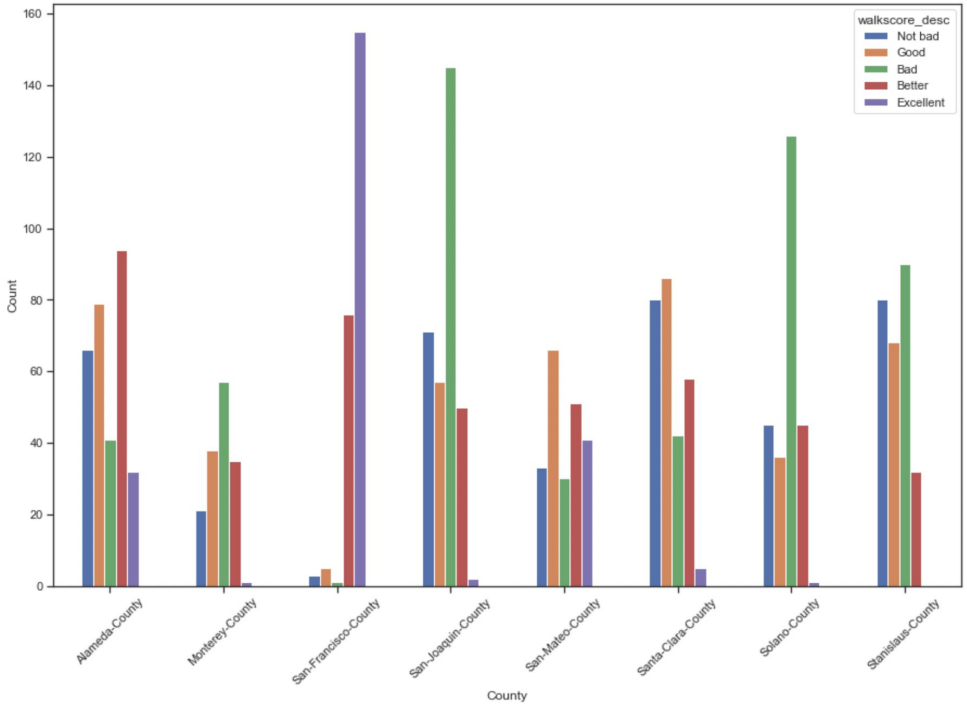
- San Francisco County has most properties that have Walk Score above 70 and has no properties with Walk Score between 0 to 40.
- Most counties have walk score in the range of 20 to 60.

Catplot of Bike Score County Wise



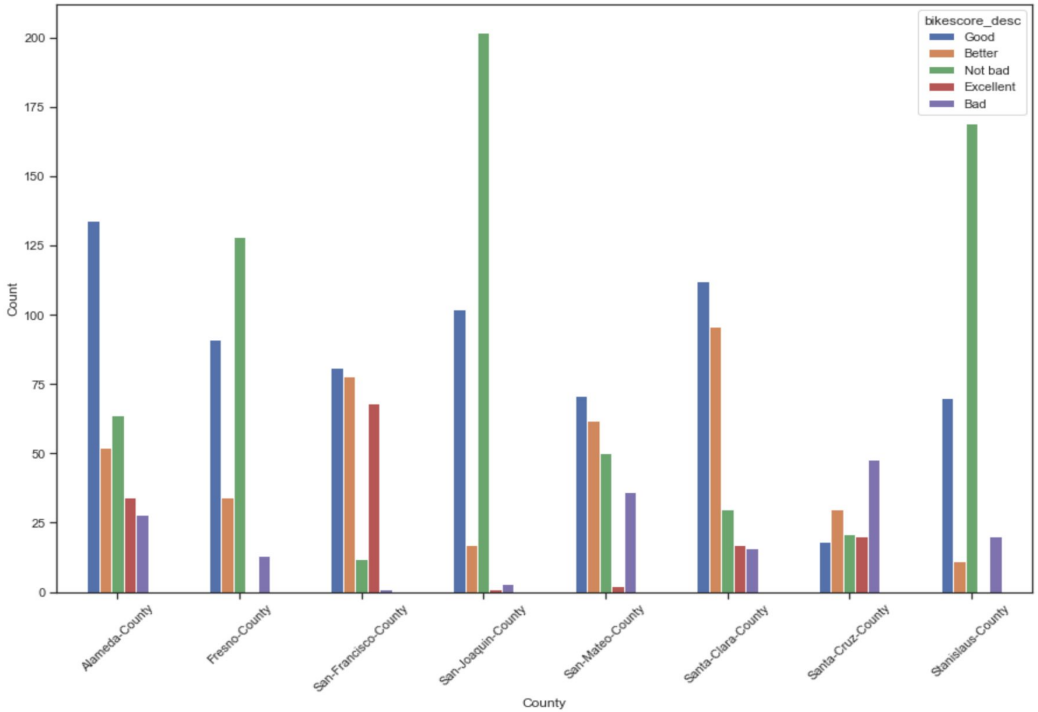
- Alpine County has no properties with Bike Score above 15.
- Colusa,Glenn, San Francisco, San Joaquin, Sutter and Yolo County have very minimal or zero bike Score under 30
- About 50% of the counties have bike score ranging from 25 to 80.

Categorizing Walk scores into ordinal categories



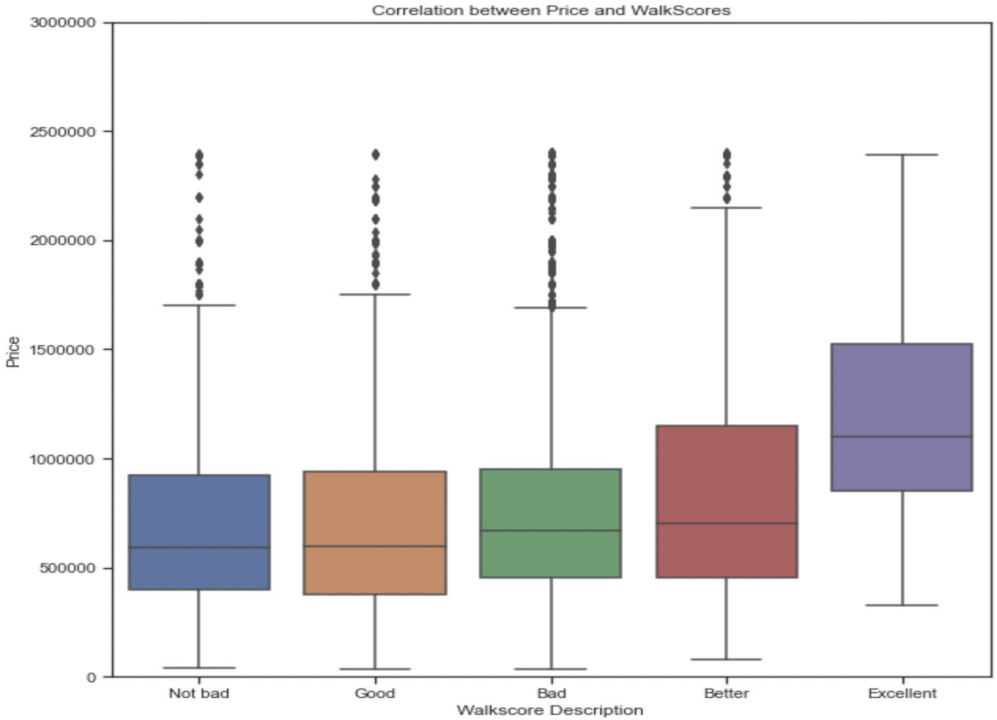
- Taking top 8 counties from Catplot with medium to good Walk Scores.
- San Francisco has an excellent Walk Score and hence most of the daily errands and activities are accomplished on foot. Shops, Malls, Restaurants, Offices etc are easily accessible in a walking distance.

Categorizing Bike scores into ordinal categories



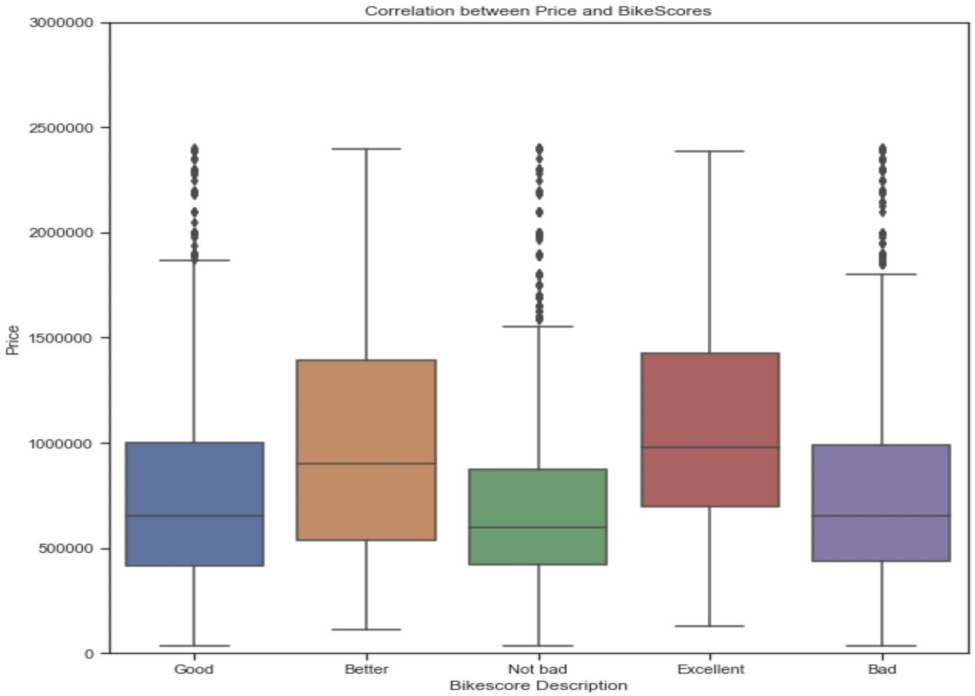
- Considering top 8 counties from the Catplot with medium to good Bike score.
- San Joaquin and Stanislaus have a very good medium range Bike Score which means most of the daily errands are accomplished on a bike.

Price range based on the Walk Score



- House price range for the counties having an excellent Walk Score is higher compared to others.
- Here, we can see that Walk score affect the pricing of houses and we notice that median value of all Categories excluding the Excellent Category is nearabout.

Price range based on the Bike Score



- It is interesting to see that though counties are rider's paradise (excellent bike score) yet, the price range is relatively largest
- We do observe that higher bike scores do affect house price ranges.

Crosstab between county & status

- Observing the Crosstab, Alameda county has the most number of new properties (New status) coming up.
- San Joaquin County has the most number of properties in Active status.

Status	Active	Active under contract	Active-reo	Back on market	Backup	Closed	Coming soon	Contingent	Contingent (no show)	Contingent (show)	Contingent - no show	Contingent - show	New	Pending	Pending (do not show)
County															
Alameda-County	66	0	1	0	0	0	1	0	0	0	0	0	237	2	1
Alpine-County	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Amador-County	80	0	0	0	0	0	0	0	0	0	0	0	1	0	0
Butte-County	165	2	0	0	0	0	0	0	0	0	0	0	0	0	0

Crosstab between county & property type

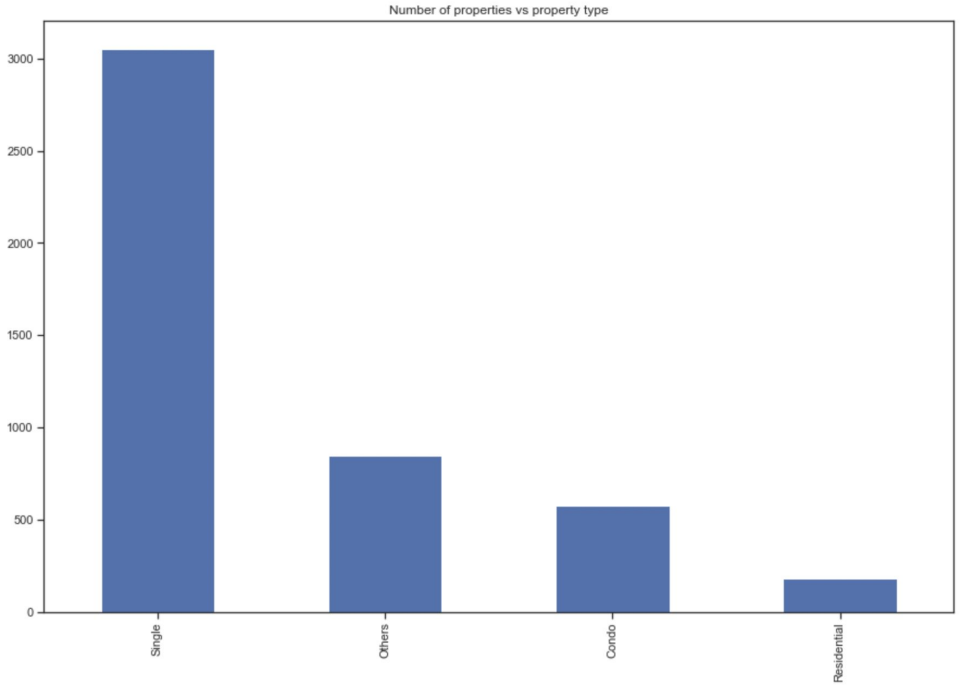
- San Joaquin County has the most number of Single Family Residential homes.
- San Francisco County has the most number of Condo Properties.
- Alpine County and Santa Cruz County have zero Single Family Residential Properties.

Property_Type	Cabin	Co-op	Commercial/residential	Condo	Condo/co-op	Detached	Double-wide mobile home	Duplex	Fourplex	Manufactured on land	...	Residential, single family	Residential, townhouse	Single family
County														
Alameda-County	0	0	0	67	0	0	2	0	0	0	...	0	0	0
Alpine-County	0	0	0	13	0	0	0	0	0	0	...	2	1	0
Amador-County	0	0	0	2	0	0	0	0	0	0	...	1	1	3
Butte-County	0	0	1	1	0	0	0	2	0	26	...	0	0	0
Calaveras-County	0	0	0	0	0	0	0	0	0	1	...	0	0	1

Categorizing property types into generic categories

- **Single** (Single family, Single family home, Single family residence, Single family residential, Single-wide mobile home)
- **Residential** (Residential-single family, Residential-townhouse, Townhouse)
- **Condo** (Cabin, Co-op, Condo, Condo/Co-Op)
- **Others**

Bar plot of Property type category vs No. of properties



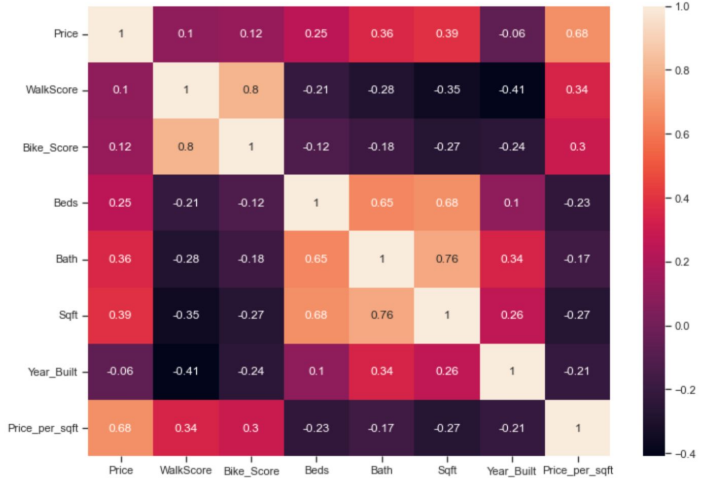
- As per the Bar plot, most of the Property types are Single.
- Residential property type seems to have least number of properties.
- Came across an article which mentions that single family homes have higher resale value and are cheaper in longer run and that supports the graph.

Prediction Analysis - Regression Model

Prediction Analysis

```
#Plotting correlation matrix to see how Price is correlated with other variables
correlation_matrix = df.corr().round(2)
sns.heatmap(data=correlation_matrix, annot=True)
# It shows price is correlated to Sqft followed by Bath and Bed.
# It also shows that years built have a negative correlation with a price which means land maintains its value over the
# and additional significant modification done on the property would not result in significant gain.
```

```
<AxesSubplot:>
```



- Plotting correlation matrix to see how Price is correlated with other attributes.
- Price is correlated more with Sqft followed by Bath and Beds.
- *Note - Price has the highest correlation with Price_per_sqft but Price_per_sqft is a calculated field based on Price itself and hence, the high correlation and will not be considered as one of the independent variables.

Prediction Analysis - Steps

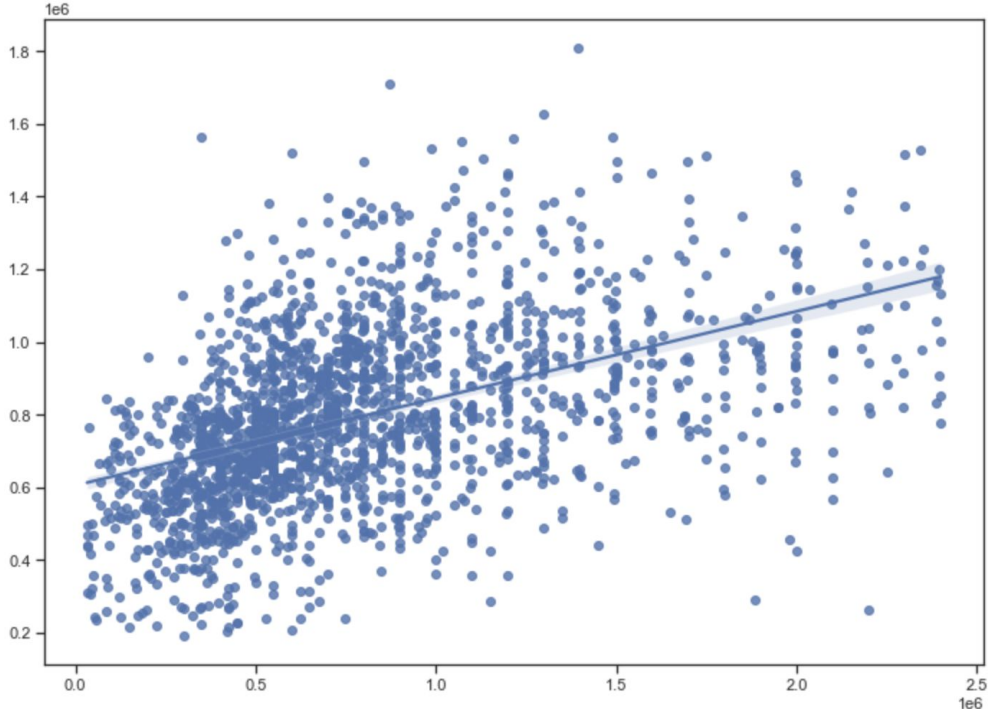
- Extracting relevant independent variables (X)- Sqft, Bed, Bath, Walk Score and Bike Score
- Price would be our dependent or target variable (Y)
- Splitting data into test and train data using sklearn
- Standardizing the attributes and fitting and transforming using StandardScaler
- Creating and fitting the linear regression model
- Comparing actual vs predicted values and calculating mean absolute error, mean squared error, root mean squared error, variance score to make inferences.

Prediction Analysis - Calculations

```
#actual vs predicted and errors
df_pred = pd.DataFrame({'Actual':Y_test,'Predicted':Y_pred})
print(df_pred.head())
from sklearn import metrics
print('Mean Absolute Error: ', metrics.mean_absolute_error(Y_test,Y_pred))
print('Mean Squared Error: ', metrics.mean_squared_error(Y_test,Y_pred))
print('Root Mean Squared Error: ', np.sqrt(metrics.mean_absolute_error(Y_test,Y_pred)))
print('Variance Score: ', metrics.explained_variance_score(Y_test,Y_pred))
```

```
      Actual     Predicted
0  2395000.0  7.802693e+05
1   550000.0  7.874840e+05
2   480000.0  5.532228e+05
3  1050000.0  1.049550e+06
4   550000.0  7.470204e+05
Mean Absolute Error:  332437.1949045364
Mean Squared Error:  191516021875.83957
Root Mean Squared Error:  576.5736682372309
Variance Score:  0.26063739879356473
```

Prediction Analysis - Regression Plot



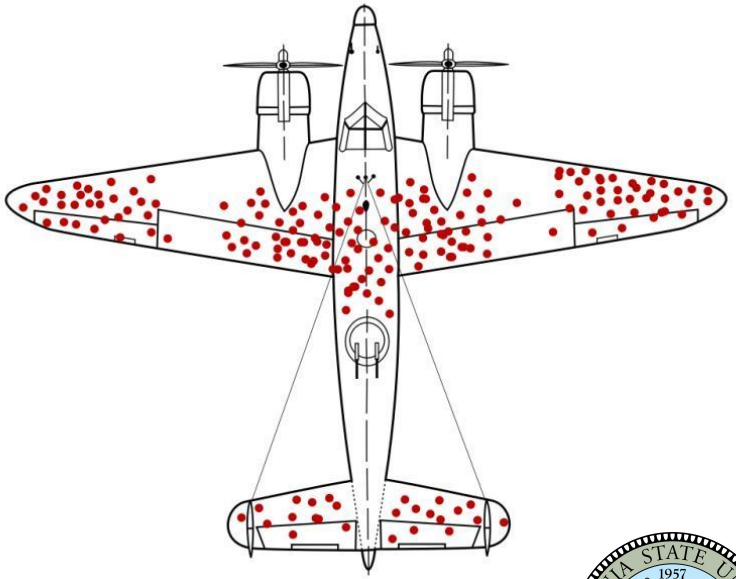
- Regression plot between actual and predicted price
- X-axis - Actual
- Y-axis - Predicted

Prediction Analysis - Inference

- The mean absolute error between the actual and predicted price is $332K
- The variance score is 0.2606 which means that model explains only 26.06% of the variance of the target variable.
- Variance score is low and it implies that the independent variables used are not enough to explain the price of the house.
- Regression plot shows that the model is not well fit and the distance between points and regression line is high (error is high)

Real World Example

- During World War II, fighter planes would come back from battle with bullet holes. The Allies found the areas that were most commonly hit by enemy fire. They sought to strengthen the most commonly damaged parts of the planes to reduce the number that was shot down.
- A mathematician, Abraham Wald, pointed out that perhaps there was another way to look at the data. Perhaps the reason certain areas of the planes weren't covered in bullet holes was that planes that were shot in those areas did not return. This insight led to the armor being re-enforced on the parts of the plane where there were no bullet holes.
- The story behind the data is arguably more important than the data itself. Or more precisely, the reason behind why we are missing certain pieces of data may be more meaningful than the data we have.



Thank you!