Datahacks Final Report

Ruidi Zhao, Junyi Zhang, Xiaotong Zeng

1. Data Cleaning

(a) Convert data types

The first step we did is trying to transform the values from string to integers or floats. As we noticed, most of the values stored in the two CSV files are in the form of string, and in order to apply some arithmetics on them, we need to first strip the extra characters in them such as "$" in most of the price columns and "," in most of the number columns.

(b) Merge Columns

We combined columns that have similar attributes in order to make the calculation easier. One example is that we merged the data for men and women of different ages and added them to two new columns called "Total Men" and "Total Women".

		Total men	Total women
Census Tract Name	Block Group		
1	1	405	496

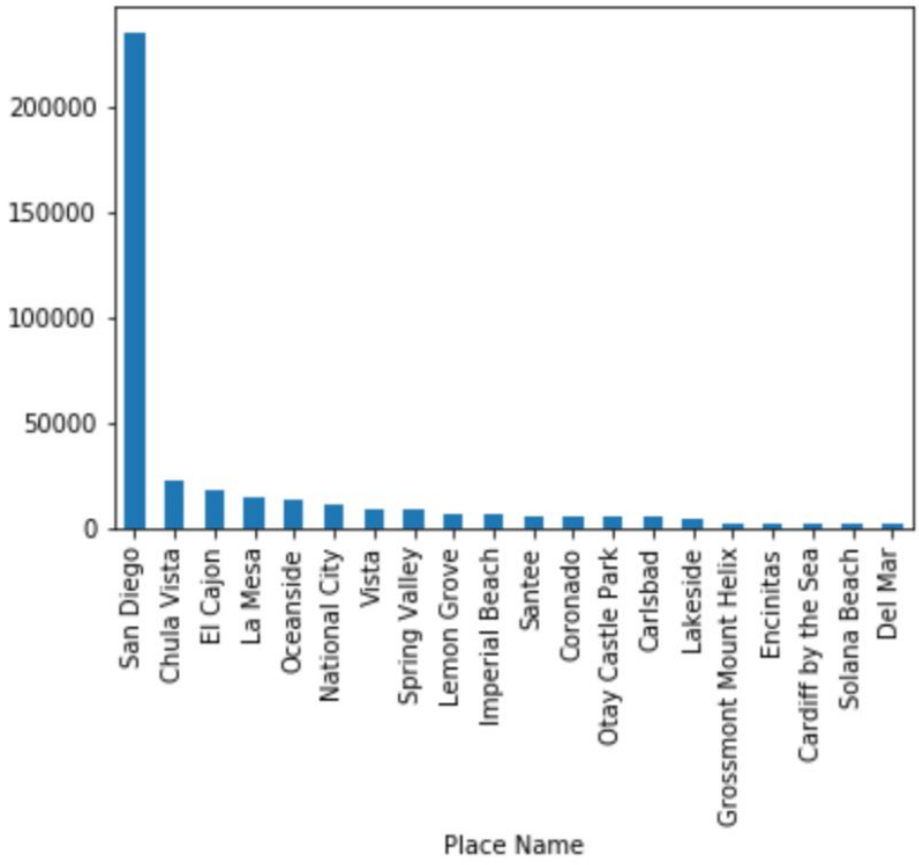
(c) *Filter "..." values*

Both datasets contain many "..." values, and we cleaned up these values based on specific questions. For example, if we are calculating the sum of a column, we would replace "..." with 0, since it won't have any influence on our final result.

For other situations, if the values we were looking for do not associate with the "...", we would delete the rows that contain "..." values.

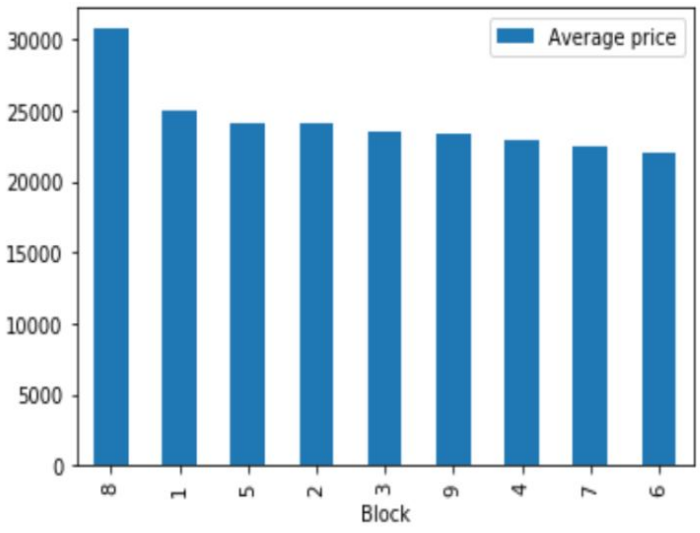
2. Visualization

(a) The following is an overall analysis of the housing distribution in the San Diego Area. According to the graphs below, San Diego has 235,346 housing units and it is the most common place that people live in. San Diego is followed by Chula Vista with 22,951 housing units, El Cajon with 17,991 housing units, and La Mesa 13,999 units.



Place Name	
San Diego	235346
Chula Vista	22951
El Cajon	17991
La Mesa	13999
Oceanside	13663
National City	11126
Vista	8651
Spring Valley	8406
Lemon Grove	6511
Imperial Beach	6043
Santee	5979
Coronado	5304
Otay Castle Park	5293
Carlsbad	5148
Lakeside	3789
Grossmont Mount Helix	2476
Encinitas	2068
Cardiff by the Sea	1963
Solana Beach	1810
Del Mar	1656

(b) Below is a bar graph that displays the average housing price distribution in San Diego. According to our data, Block 8 in the city of San Diego has the highest average price among all blocks. The average price of all houses in San Diego is around $24,029.



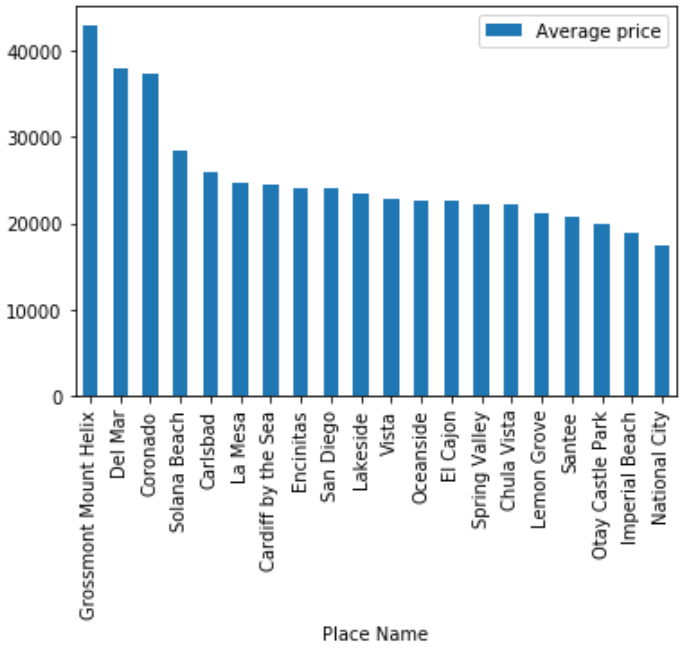
```
In [2]: average = housing["Total price"].sum()/housing["Total units"].sum()

In [3]: average
Out[3]: 24029.803669852485
```

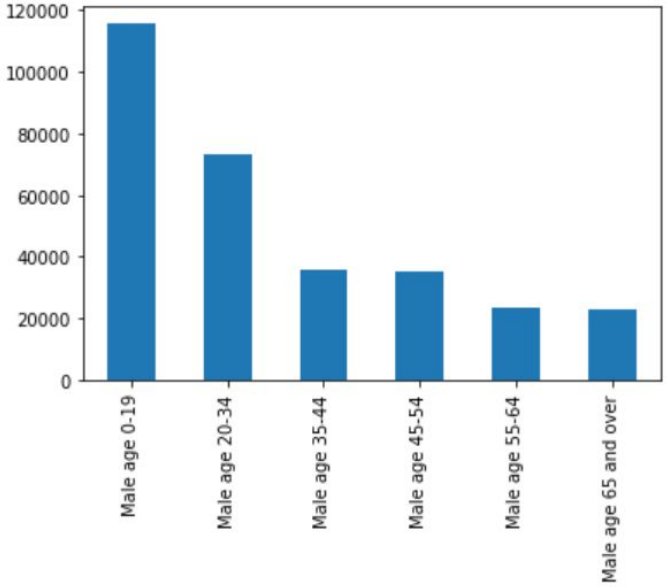
(c) Based on the owner-occupied average value, Grossmont Mount Helix has the highest average housing price of $42,034, followed by Del Mar with $38,003 and Coronado with $37,452.

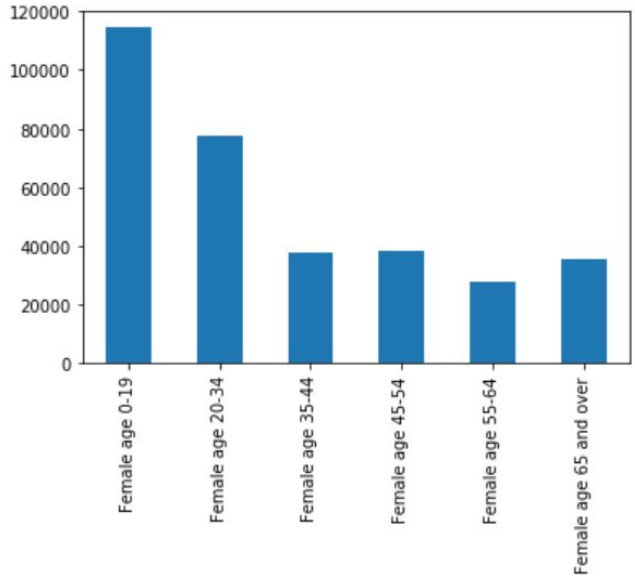
```
In [65]: housing
Out[65]:
```

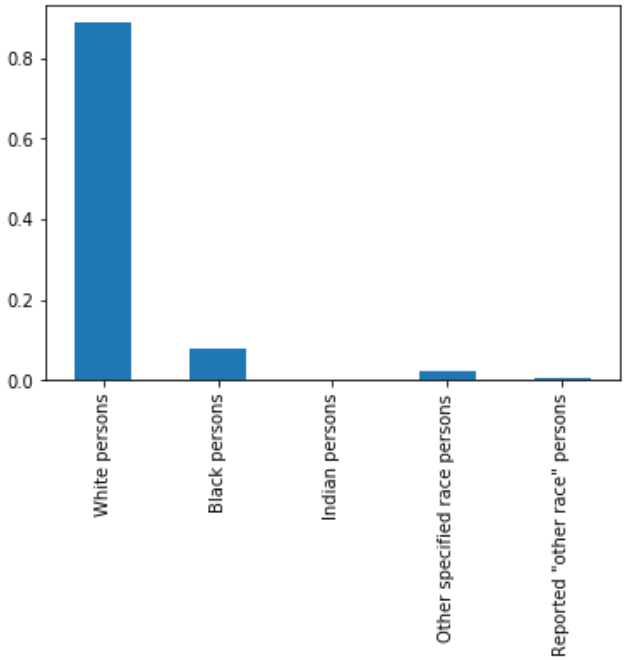
Place Name	Average price
Grossmont Mount Helix	43034.288718
Del Mar	38003.456522
Coronado	37452.887819
Solana Beach	28385.477348
Carlsbad	25972.937755
La Mesa	24768.105744
Cardiff by the Sea	24526.130922
Encinitas	24040.106867
San Diego	24029.803670
Lakeside	23433.555819
Vista	22895.847272
Oceanside	22681.846655
El Cajon	22518.547502
Spring Valley	22285.298584
Chula Vista	22248.422236
Lemon Grove	21133.077112
Santee	20831.253429
Otay Castle Park	19818.944423
Imperial Beach	18963.264647
National City	17495.692498

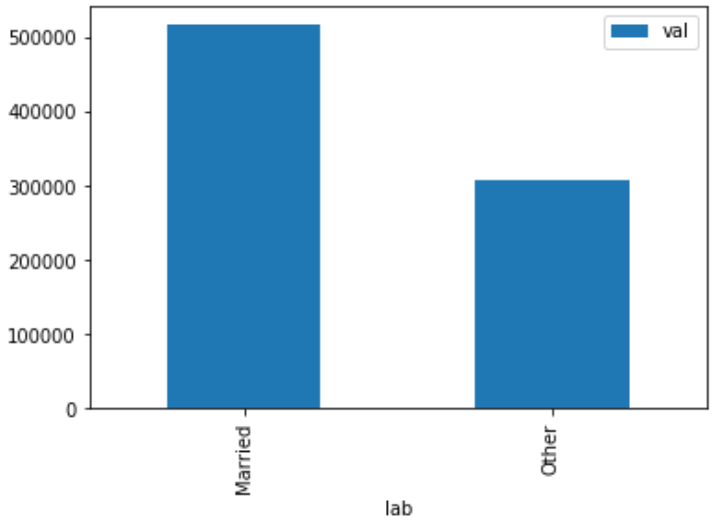


(d) The following graphs demonstrate the gender distribution, racial distribution (in percentages), marital status distribution in the area of San Diego.









3. Machine learning

(a) First, we define the average of housing price as the living condition in a specific area. After looking through the entire dataset and previous graphs we plotted above, we found that the gender-ratio (male:female), married status (married:not married), and race ratio (caucasian : others) are strongly correlated with the average price of housing. Therefore, we merge these three parameters in terms of track number and block number. With the help of package Sklearn, we ran the OLS method to regress these three independent variables on average housing price and got the coefficients for each. By using the t-test, we also found that the coefficients we obtained are all statistically significant with mean squared error 0.63.

```
In [443]:
t_stat_men_ratio = reg.coef_[0]/std_Men
t_stat_married_ratio = reg.coef_[1]/std_Married
t_stat_White_ratio = reg.coef_[2]/std_White
print('men_ratio_test:', t_stat_men_ratio)
print('married_ratio_test:', t_stat_married_ratio)
print('white_ratio_test:', t_stat_White_ratio)

men_ratio_test: 21.697164751302196
married_ratio_test: -5.999439533546715
white_ratio_test: 6.638583750988894
```

```
print('Mean Absolute Error:', metrics.mean_absolute_error(Merged["Average price"], Merged["Prediction"]))
print('Mean Squared Error:', metrics.mean_squared_error(Merged["Average price"], Merged["Prediction"]))
print('Root Mean Squared Error:', np.sqrt(metrics.mean_squared_error(Merged["Average price"], Merged["Prediction"])))

Mean Absolute Error: 0.5860482770045579
Mean Squared Error: 0.6299766051909734
Root Mean Squared Error: 0.7937106558381167
```

Our linear regression model is:

Average price = Men_ratio * 11251.63 + Married_ratio * -6371.86 + White_ratio * 15639.66 + 7509.37

Census Tract Name	Block Group	Men Ratio	Total women	Married Ratio	White ratio	Average price
1	1	0.449501	496	0.585652	0.980022	32298.0
	2	0.439239	383	0.570874	0.985359	29485.0
	3	0.468045	283	0.616708	0.994361	37151.0
	4	0.467933	224	0.617978	0.985748	44712.0
	5	0.433538	277	0.636591	0.987730	33229.0
	6	0.433735	282	0.626536	0.995984	40398.0
2	1	0.453789	591	0.556314	0.995379	29360.0
	2	0.463970	1153	0.538987	0.991167	22080.0
	3	0.505935	666	0.564947	0.985163	27171.0
	4	0.481250	498	0.541667	0.987500	23330.0

4. Conclusion

After we clean the data and visualize the patterns, we found that gender ratio, race ratio, and marital status are strongly correlated with the average price of housing, which indicated the living condition in a given area. However, we found that the gender ratio and race ratio have a positive effect while marital status ratio has a negative effect on the dependent variable. According to our predictive model, if we have more men and the Caucasian population in an area, the average price of housing in that area will be expected to be higher than average. On the contrary, if an area has many couples with other factors as constant, the average price of housing will be lower. However, although all of the coefficients are statistically significant, the R^2 is only 0.174, which is not too high. This flaw can be explained by the fact that we might miss some of factors which are also correlated with the dependent variable.