# Case Study: Regression

Mid-bootcamp project - Regression

# Index

## SQL

- Main challenges
- Achievements

## TABLEAU

- Main Challenges
- Dashboards

## PYTHON

- Machine Learning Process
- ML Visualization
- Achievements
- Main Challenges

Mid-bootcamp project - Regression

# SQL

## Main Challenges

- Data Extraction
- Query Comprehension

## Achievements

- 13/14  Solved Questions

Q12: properties whose prices are twice more than the average of all the properties in the database:

```
select id
from house_price_data
Where price > (Select avg(price)*2
                from house_price_data);
```

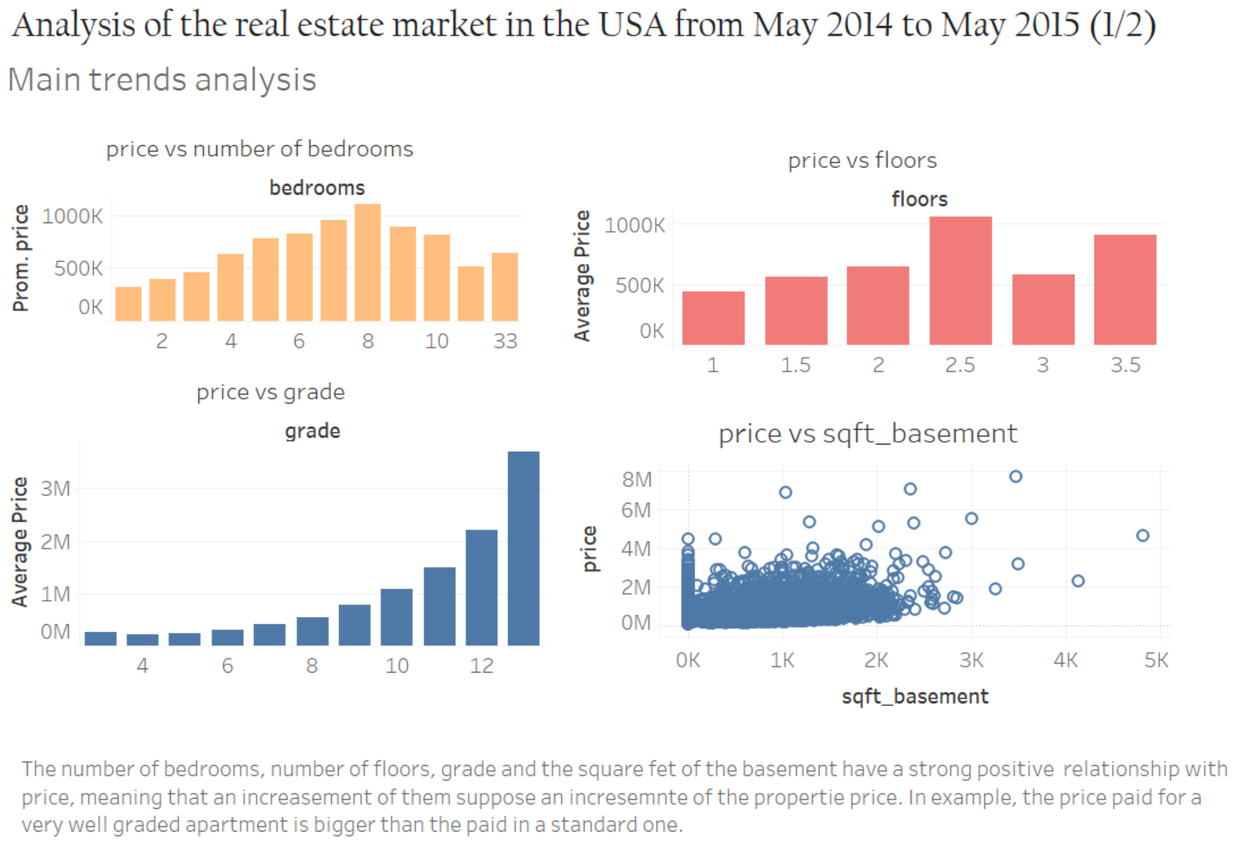
| id |
|---|
| 2147483647 |
| 2147483647 |
| 822039084 |
| 1802000060 |
| 2147483647 |
| 2147483647 |

# Tableau

## Main Challenges

- Overcomplicating
- Extract self conclusions

## Dashboard

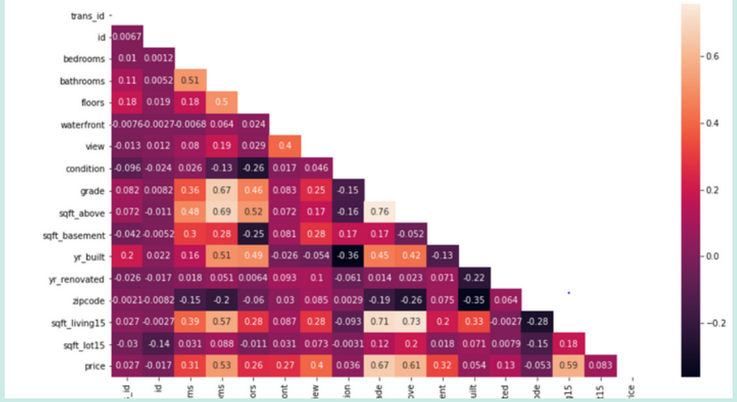


# Python

## DATA CLEANING, WRANGLING AND PREPORCESSING

For having a better visualization od the correlation between variables we plot a head map

```
mask=np.zeros_like(data_1.corr())
mask[np.triu_indices_from(mask)]=True
fig, ax=plt.subplots(figsize=(14, 8))
ax=sns.heatmap(data_1.corr(), mask=mask, annot=True)
plt.show()
```




## MACHINE LEARNING MODEL

Import the model and fit it with our train data:

```
from sklearn.linear_model import LinearRegression
regressor = LinearRegression()
regressor.fit(X_train, y_train)
```

### APPLYING THE MODEL: OLS MODEL

Import the library and fit the model

```
import statsmodels.api as sm

X_train = sm.add_constant(X_train)

model = sm.OLS(y_train, X_train).fit()
predictions = model.predict(X_train)

print_model = model.summary()
print(print_model)
```

## MODEL EVALUATION AND MAIN FINDINGS

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.525
Model:                            OLS   Adj. R-squared:                  0.525
Method:                 Least Squares   F-statistic:                     3790.
Date:                Thu, 19 Nov 2020   Prob (F-statistic):               0.00
Time:                        21:15:00   Log-Likelihood:            -2.3585e+05
No. Observations:               17125   AIC:                         4.717e+05
Df Residuals:                   17119   BIC:                         4.718e+05
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
```

# Python

Mid-bootcamp project - Regression

## ACHIEVEMENTS

- Applied 3 ML Models (52% Accuracy)
- Learned different techniques

## MAIN CHALLENGES

- Apply the wrong Model
- Follow carefully the steps
- Planification

## LEARNINGS

- Plan better
- Apply Mulltpile ML Models
- Adapt

# Thank you!