

**Northeastern University**

**Intermediate Analytics**

**ALY6015**

**Module – 6 Final Project Report**

**Submitted to: Prof. Roy Wada**

**Fnu Abhishek,**

**Kyla Araujo,**

**Balram Marni,**

**Ayushi Ajay Walia**

**Introduction**

Boston's housing market is known for being competitive and expensive, with high demand and limited supply driving up prices. The city's real estate market is heavily influenced by its many universities and colleges, which attract a large number of students and faculty members to the area each year.

Despite the high cost of living in Boston, the city remains a popular destination for people looking for a vibrant urban lifestyle, excellent job opportunities, and access to world-class education and healthcare.

We have used Boston Assessment 2015 dataset which contains one hundred sixty-eight thousand and one hundred and sixteen records and it has fifty-six columns.

**Analysis**

1. We have provided the *Statistical Summary Chart* of the continuous variables based on Boston Housing Dataset:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **AV\_LAND** | **AV\_BLDG** | **GROSS\_TAX** | **LAND\_SF** | **GROSS\_AREA** | **LIVING\_AREA** |
| **Min.:**  100 | **Min.:** 3.440e+02 | **Min.:** 0 | **Min.:** 0 | **Min.:** 0 | **Min.:** 0 |
| **1st Qu.:** 111800 | **1st Qu.:** 1.799e+05 | **1st Qu.:** 3539 | **1st Qu.:** 4087 | **1st Qu.:** 2436 | **1st Qu.:** 1421 |
| **Median:** 142900 | **Median :** 2.418e+05 | **Median:** 4592 | **Median:** 5286 | **Median:** 3294 | **Median:** 2006 |
| **Mean:** 1818254 | **Mean :** 3.461e+06 | **Mean:** 26822 | **Mean:** 211619 | **Mean:** 10611 | **Mean:** 8520 |
| **3rd Qu.:** 193800 | **3rd Qu.:** 3.660e+05 | **3rd Qu.:** 6670 | **3rd Qu.:** 7252 | **3rd Qu.:** 5051 | **3rd Qu.:** 3166 |
| **Max.:** 372028900 | **Max.:** 1.245e+09 | **Max.:** 20110396 | **Max.:** 101596835 | **Max.:** 1976650 | **Max.:** 1940476 |

***Table 1.1: Statistical Summary Chart***

1. We have used bar chart, we have shown the average building cost varying on different kitchen types such as **F, O, N,** and **P**:
2. **F:** A *full eat-in* kitchen is a type of kitchen that includes a designated space for dining within the kitchen area. The average building cost in this category is **$492.761 K**.
3. **N:** *None* category has the **highest** average building cost of **$1.51** **million** approximately.
4. **O:** The layout of a *one-person* kitchen usually includes essential elements such as a sink, stove, refrigerator, and storage cabinets, all arranged in a way that maximizes the available space. The average building cost in this category is **$513.10 K**.
5. **P:** A *pull/alcove* kitchen is a type of kitchen design that is characterized by a layout where the kitchen area is set back or recessed from the surrounding walls or living space. It has the **lowest** average building cost of **$342 K** approximately.

Chart, bar chart

Description automatically generated

***Figure1.1: Bar Chart of average building cost varying on different kitchen types***

1. In the below bar chart, we have shown the variations of average **building cost** in every decade from 1920 till 2010. We have analyzed that average building cost has the **maximum** value of **$7.38 million** approximately in the decade **1970** whereas the **minimum** value of average building cost was recorded as **$729.747 K** in the decade **1930**.

Chart, bar chart

Description automatically generated

***Figure1.2: Bar chart of building expenses across several decades***

1. In the below bar chart, we have shown the variations of average land cost in every decade from 1920 till 2010. We have analyzed that average land cost has the **maximum** value of **$3.4 million** approximately in the decade **1960** whereas the minimum value of average land cost was recorded as **$326.176 K** in the decade **1930**.

Chart, bar chart

Description automatically generated

***Figure1.3: Bar chart to depict the land cost across several decades***

1. In the below bar chart, we have shown the variations of average *gross tax* in every decade from 1920 till 2010. We have analyzed that average gross tax has the **maximum** value of **$86.6 K** approximately in the decade **1970** whereas the **minimum** value of average gross tax was recorded as **$7,303** in the decade **1950**.

Chart, bar chart, histogram

Description automatically generated

***Figure1.4: Bar Chart for displaying gross tax across several decades***

1. In the below histogram chart, we have shown the total number of houses available for purchase in every decade from 1920 till 2010. The **highest** number (**18,778**) of houses were available in the decade **1920** whereas the **lowest** number (**1939**) of houses were available in the decade **2010**.

Chart, bar chart, histogram

Description automatically generated

***Figure1.5: Bar Chart of total housing for rent in every decade***

1. In the pie chart, we have shown total percentage of each structure class available in the given dataset:
2. **Struct Steel (A)**: **0.6%**
3. **Reinforced Concrete (B)**: **1.25%**
4. **Brick/Concrete (C): 0.12%**
5. **Wood/Frame (D): 4.72%**
6. **Metal (E): 0.12%**
7. **Residential (R): 79.28%**

Chart, pie chart

Description automatically generated

***Figure1.6: Pie Chart of structure class available in given dataset***

1. In this correlation matrix, we find out the different relations between Land Cost (av\_land), Building Cost (av\_bldg ),Gross tax (Gross\_Tax) and Land Surface Area (Land\_SF), Living Area (LIVING\_AREA), Gross Area (Gross\_Area). We have observed that the correlation between *Gross Area* and *Living Area* has the **highest** value **0.98888374** whereas correlation between *Gross Tax* and *Land Surface Area* has the **lowest** value **0.002140853**.

***Figure1.7: Corrplot*** Chart, scatter chart

Description automatically generated

1. In the below scatter plot, we have shown the variations of the Land Cost as per the Building cost. We have also analyzed the linear equation of line as:

*Y = 4 \* 105 + 1.5 \*X*

Where:

Y = Building Cost,

X = Land Cost

The **slope** of the linear equation is **non-zero** which means the predictor variable is the weak entity and also **R2** value is **0.4597596**. By observing these two conditions we can conclude that these variables are not linearly dependent on each other.

Chart, scatter chart

Description automatically generated

***Figure1.8: Scatter plot of Building cost v/s Land cost***

1. In the below scatter plot, we have shown the variations of the Land Cost as per the Land Surface Area. We have also analyzed the linear equation of line as:

*Y = -61000 + 0.2 \* X*

Where:

Y = Land Surface Area,

X = Land Cost

The **slope** of the linear equation is **non-zero** which means the predictor variable is the weak entity and **R2** value is **0.654135**. By observing these two conditions we can conclude that these variables are not linearly dependent on each other.

Chart, scatter chart

Description automatically generated

***Figure1.9: Scatter plot of Surface area and Land cost***

1. In the below scatter plot, we have shown the variations of the Gross Tax as per the Living Area. We have also analyzed the linear equation of line as:

*Y = 2300 + 0.1 \* X*

Where:

Y = Living Area,

X = Gross Tax

The **slope** of the linear equation is **non-zero** which means the predictor variable is the weak entity and also **R2** value is **0.4544352**. By observing these two conditions we can conclude that these variables are not linearly dependent on each other.

Chart, scatter chart

Description automatically generated

***Figure1.10: Scatter plot for average Living area and gross tax***

1. We have done linear regression analysis between *Land Cost* and *Building Cost* (AV\_BLDG) along with *Land Surface Area* (LAND\_SF). We have also analyzed the linear equation of line as:

*Y = 157424.6693* *+ 0.221557566\* X1 + 2.697764655 \* X2*

Where:

Y = Land Cost,

X1 = Gross Tax,

X2 = Land Surface Area

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Coefficients** | **R-Square** | **AIC** |
| **(Intercept)** | 157424.6693 | 0.859544618 | 5402433.9 |
| **AV\_BLDG** | 0.221557566 | 0.859544618 | 5402433.9 |
| **LAND\_SF** | 2.697764655 | 0.859544618 | 5402433.9 |

***Table1.2: Linear regression analysis between Land Cost and Building Cost***

1. We have done another linear regression analysis between *Gross Tax* and *Living Area* (LIVING\_AREA). We have also analyzed the linear equation of line as:

*Y = 2282.98466 + 0.10157837\* X*

Where:

Y = Gross Tax,

X = Living Area

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Coefficients** | **R-Squared** | **AIC** |
| **(Intercept)** | 2282.98466 | 0.4544352 | 3716795.53 |
| **GROSS\_TAX** | 0.10157837 | 0.4544352 | 3716795.53 |

***Table1.3: Linear regression analysis between Gross Tax and Living Area***

**Conclusion**

1. From the analysis of part 2, we can conclude that Kitchen Type does not highly correlated with Building Cost.
2. From the analysis of parts 3, 4 and 5, we can conclude that in the 1960s and 1970s, the city faced a decline in population and economic activity, leading to a decrease in property values and an increase in vacant properties whereas in the 1980s and 1990s saw a resurgence in the Boston real estate market, driven in part by the city's booming economy and a wave of new development.
3. From the analysis of part 6, we can conclude that in the 1920s and 1930s, Boston experienced a period of urban renewal and redevelopment, which included the construction of new housing and commercial buildings.
4. From the analysis of parts 9, 10 and 11, we can conclude that the best suited linear regression model can be built between Land Cost and Land Surface Area.
5. From the analysis of parts 12 and 13, we can conclude that the regression model based on *Land Cost* and *Building Cost* (AV\_BLDG) along with *Land Surface Area* (LAND\_SF) **better** than the linear regression model based on *Gross Tax* and *Living Area* (LIVING\_AREA) as per the **R2** and **AIC** (Akaike Information Criterion) value.