**A Data Analytics Methodology for Analyzing Real Estate Brokerage Markets: A Case Study of Dubai**

Ahmed Saif, Maged Al Hashemi, Mohammed Aleissaee, Abdelaziz Almansoori

**Abstract**

**This report provides an analytics methodology and insights about Real Estate brokerage market. First, we develop our methodology that integrates various statistical and analytical methods to construct and interpret the graphs in the analysis part. Then, the applicability of the methodology is illustrated with a case study, where data for top 50 real estate brokerage firms in Dubai are analyzed. In particular, graphs visualizations provide a better understanding of the real estate broker, as demonstrated in the case study.**

**Keywords**

Real Estate Brokers, Analytics Workflow, Graph Analytics, Multi-Dimensional Scaling (MDS)

**Introduction**

Real Estate is one of the primary type of investment for investor, andtypically considered lower risk with stable consistent returns on investment. Much of the real estate transactions are facilitated through real estate brokerage firms. There are many companies and offices in the real estate sector in every major city, including Dubai, which is the focus of our case study. The number of brokerage companies in the real estate sector is increasing each year, yet there are only limited data analytics studies that investigate how the brokerage companies in the real estate market can be benchmarked. In this study, we analyze the brokerage companies in the real estate market in Dubai, using public data by Dubai Land Department [1], the government body in Dubai that regulates, monitors, and governs the real estate sector in Dubai. To analyze the data at hand, it was necessary to develop a customized data workflow that usees various statistical and analytical methods, consisting of two sample comparison of means (Student's t test), visual analytics, hierarchical clustering, k-Means clustering, MDS (Multi-dimensional Scaling), outlier analysis, and linear regression.

**Literature**

Our analysis is based on Real Estate Brokerage Markets from the website of Dubai Government.

Obinna and Udo (2022) [2] investigate making better decisions based on the office broker's data by applying data analytics and statistical visualization tools to demonstrate insights. The authors apply clustering and linear regression techniques to provide insights based on data from over 62 PropTech firms in Nigeria, providing customers, managers, and agents with a better knowledge of the real estate brokerage industry.

Razen et al. (2014) [3] firstly outline the steps taken to anticipate many real estate market attributes using big data. Then, the authors present several statistical models, in particular multilevel structured additive regression models, to identify the statistically significant attributes that affect the valuation for the real estate.

**Methodology**

* In this research study, we developed an analytics methodology for analyzing brokerage companies in the real estate sector.

A screenshot of a computer

Description automatically generated with medium confidence

**Figure 1.** The data analytics methodology developed and applied in the study.

* Figure 1 displays the analytics methodology as a workflow in Orange software (<https://orangedatamining.com>), with a variety of different statistical and analytical methods systematically applied. For example, scatterplots explain the relationships between two different variables and linear regression explains multiple variables affect multiple factors of response.

The statistical and analytical methods integrated into our methodology are as follows:

1. Scatter plot
2. Hierarchical clustering
3. k-Means clustering
4. MDS (Multi-dimensional Scaling)
5. Box plot and strip plot
6. Two sample comparison of means (Student's t test)
7. Outlier analysis
8. Linear regression

**Analysis**

As the source dataset, we extracted data for Dubai licensed real estate brokers,

* <https://dubailand.gov.ae/en/eservices/licensed-real-estate-brokers/licensed-real-estate-brokers-list/#/> [1]

obtained from the website of Dubai Land Authority

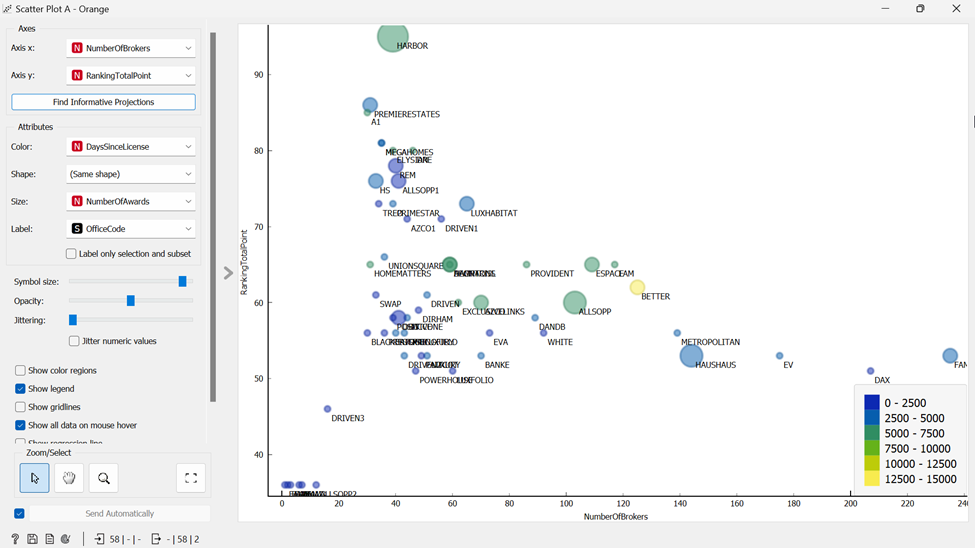
* <https://dubailand.gov.ae/>

After collecting and cleaning the data, we constructed a data analytics methodology and we conducted the analysis using Orange software.

The data attributes are provided in Table 1. Dark red text denotes the primary numerical target attribute, namely RandkingTotalPoint. The eight red colored attributes are points corresponding to different dimensions of performance. Attributes shown in green cells are derived attributes, meaning that they were derived from the original dataset through calculations. The gray attributes are not used in the analysis.

**Table 1.** Data attributes for the analyzed data.

|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute** | **Data Type** | **Description** | **IsUsed** |
| OfficeCode (Key Attribute) | Text | Custom created office code as ID attribute (to be used instead of OfficeNumber) | Yes |
| Name | Text | Full name of the office | Yes |
| TotalBrokersAllBranches | Numeric | Number of brokers in the whole company, including all offices of that company | Yes |
| NumberOfBrokers | Numeric | Number of brokers in this particular office (not the whole company) | Yes |
| RankingTotalPoint | Numeric | Total points for an office, calculated as the sum of Points attributes (out of 100) | Yes |
| PointsNumberofTransactions | Numeric | Points for number of transactions | Yes |
| PointsTransactionTotalWorth | Numeric | Points for transactions total worth | Yes |
| PointsLegalNotice | Numeric | Points for legal notice | Yes |
| PointsComplianceWithLaws | Numeric | Points for compliance with laws | Yes |
| PointsRealEstateExperience | Numeric | Points for real estate experience | Yes |
| PointsNumberOfBranches | Numeric | Points for number of branches | Yes |
| PointsLocalization | Numeric | Points for localization | Yes |
| PointsInitiatives | Numeric | Points for initivaites | Yes |
| NumberOfAwards | Numeric | Number of awards that the office has received | Yes |
| HasEmailDomain | Binary | 1 if the office has email domain of its own, 0 otherwise | Yes |
| IsMobileNumber | Binary | 1 if the office has a mobile number listed as contact, 0 otherwise | Yes |
| LicenseIssueYear | Numeric | Year in which license was issued by the Dubai Land Department | Yes |
| DaysSinceLicense | Numeric | Number of days since license was issued by the Dubai Land Department | Yes |
| *OfficeNumber* | *ID* | *Provided by the Dubai Land Authority, unique OfficeNumber for each Office* | *No* |
| *Email* | *Text* | *Contact email, provided by the Dubai Land Authority* | *No* |
| *Phone* | *Text* | *Contact phone number, provided by the Dubai Land Authority* | *No* |
| *LicenseIssueDate* | *Date* | *Date the license was issued, in date format* | *No* |
| *LicenseExpiryDate* | *Date* | *Date the license will expire, in date format* | *No* |
| *Award1* | *Text* | *Description of Award 1 received by the office* | *No* |
| *Award2* | *Text* | *Description of Award 2 received by the office* | *No* |
| *Award3* | *Text* | *Description of Award 3 received by the office* | *No* |
| *Award1Year* | *Numeric* | *Year of Award 1 received by the office* | *No* |
| *Award2Year* | *Numeric* | *Year of Award 2 received by the office* | *No* |
| *Award3Year* | *Numeric* | *Year of Award 3 received by the office* | *No* |
| *LicenseIssueMonth* | *Numeric* | *Month in which license was issued by the Dubai Land Department* | *No* |
| *LicenseIssueDay* | *Numeric* | *Day in which license was issued by the Dubai Land Department* | *No* |
| *LicenseIssueDateFormat* | *Date* | *Date the license was issued, in date format* | *No* |



**Figure 2.** Scatter plot analysis of multiple attributes at once.

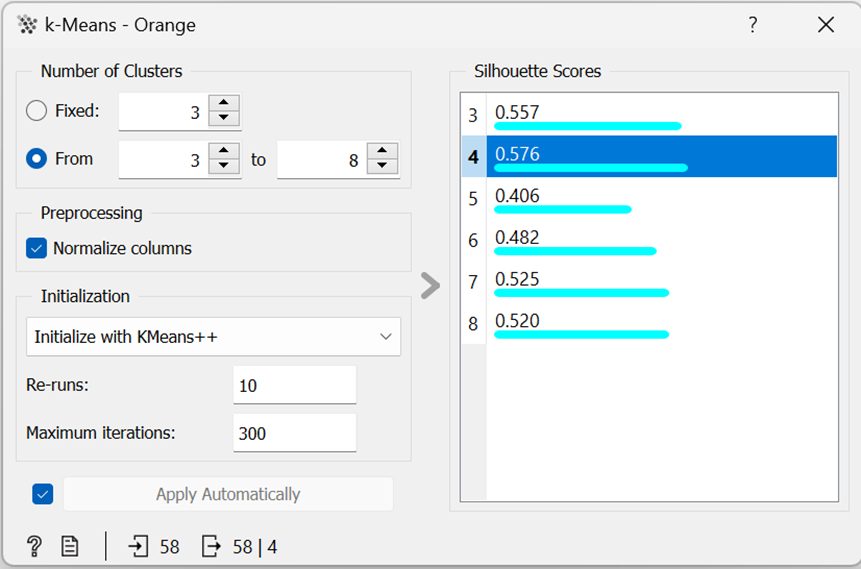
* In Figure 2, we can see that we have a scatterplot with variable “NumberOfBrokers” (Number of Brokers) as our x-axis meanwhile “RankingTotalPoint”(Ranking total point) as our y-axis. In addition, we denoted different attributes such as color “DaysSinceLicense” with “NumberOfAwards” through size. It is essential to highlight that we used “OfficeCode” as a label too.
* The scatterplot illustrates that there is no single consistent pattern, such as a positive or negative relationship between the two variables. However, this scatterplot can still yield a highly beneficial insight: A real estate buyer would prefer high values of “RankingTotalPoint” and “NumberOfAwards”. It is possible to identify such brokers from this scatterplot as the larger circles on the upper region of the plot. Furthermore, analyzing the sector, one would especially be interested in identifying younger companies with fewer employees but with high level of success. These companies are denoted by larger circles in the upper left region, as well as darker colors and towards the left side of the plot.

A picture containing text, screenshot, display, software

Description automatically generated

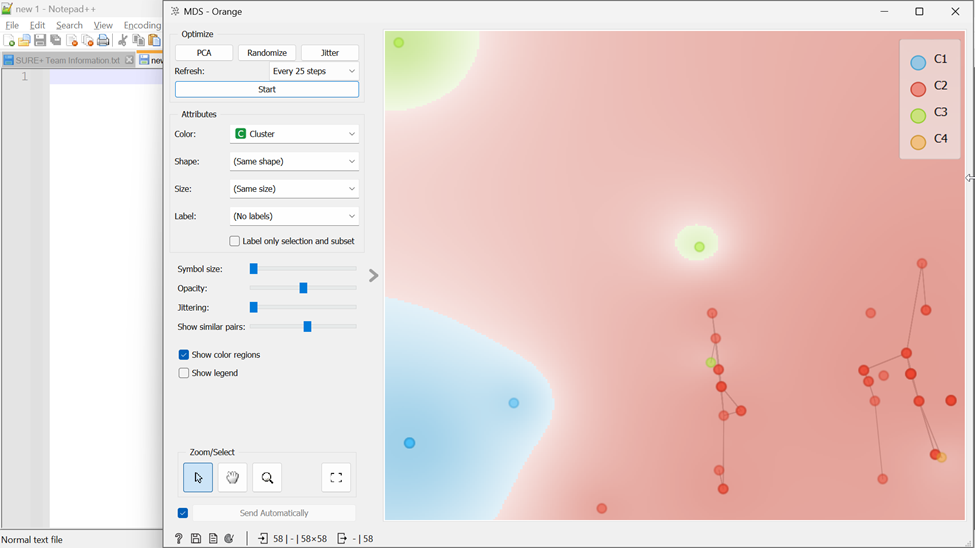
**Figure 3.** Hierarchical clustering results.

* Figure 3 shows the application of the Hierarchal clustering method with no pruning with “office code” as annotation. The hierarchal clustering allows to identify the companies that are like each other. For example, on the upper right of the figure, one can observe that HAMPTONS, AEONTRISL, PROVIDENT , FAM, and ESPACE are grouped together, meaning that these companies are similar with respect to the attributes used in clustering.



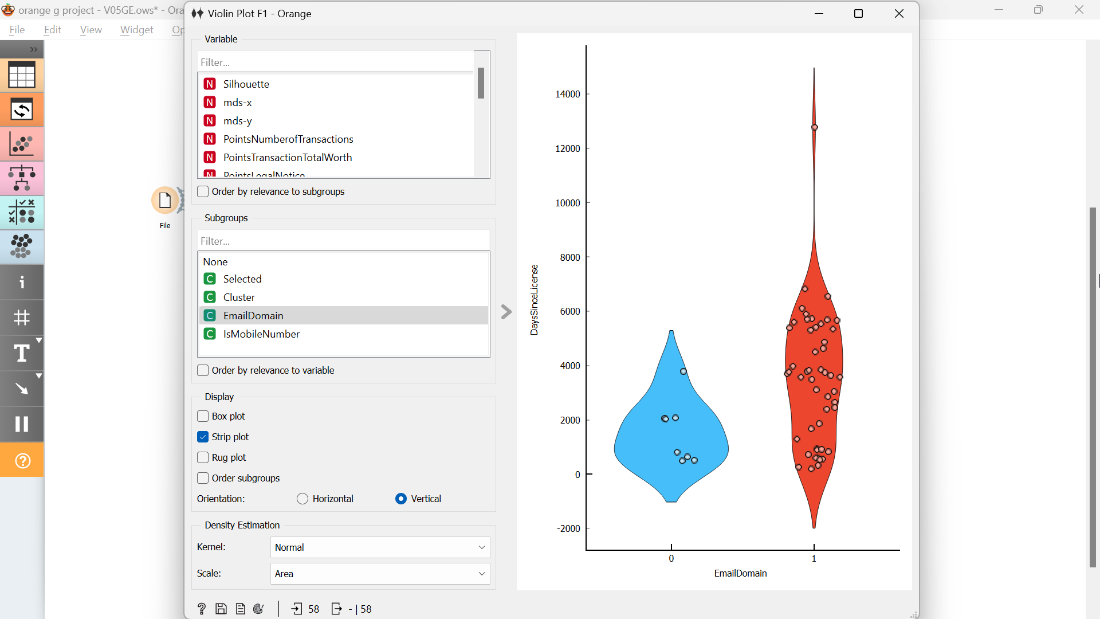
**Figure 4.** Parameters for k-Means clustering.

* The next analysis is k-means clustering, whose results highly depend on the selection of the number of clusters. Figure 4 displays the Silhouette scores for different values for k. Silhouette score is highest for k=4, thus we selected k=4 as the number of clusters. Other parameters for pre-processing are provided in Figure 4.



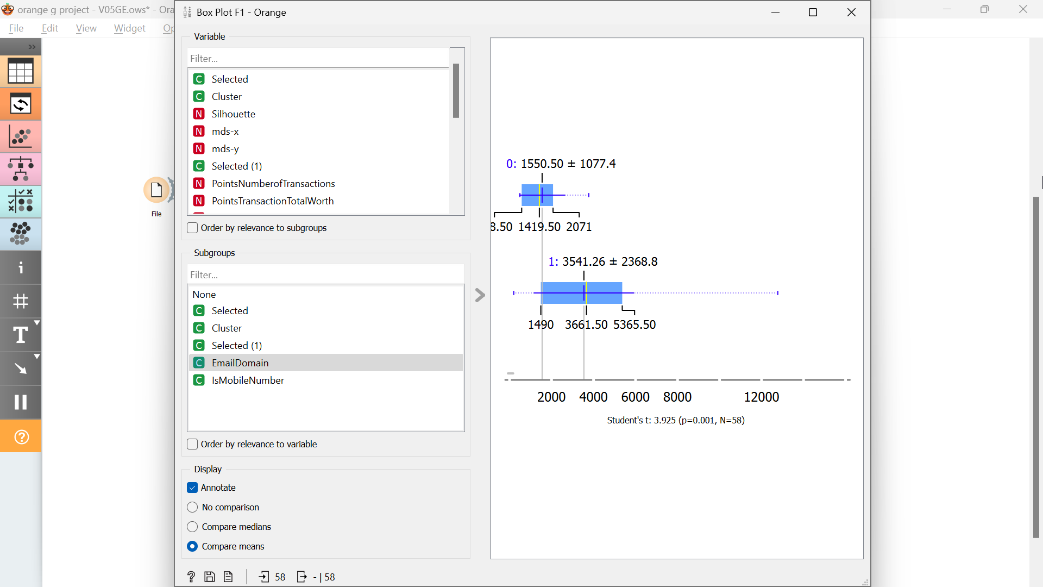
**Figure 5.** Results of multi-dimensional scaling (MDS).

* After applying k-means clustering, each observation is mapped to a cluster. A reasonable successive analysis is understanding how these observations in different clusters are related. To this end, we applied the multi-dimensional scaling (MDS) method, which maps multi-dimensional data onto two-dimensional Cartesian plane.
* In the MDS, “Show similar pairs” option is selected, which automatically turns the MDS results into a graph. In other words, MDS analysis also counts as the application of graph analytics for this domain and dataset.
* Figure 5 shows a zoomed region within the results of MDS, with color regions displayed. Color denotes the cluster Id.
* In this figure, we focused on cluster 2, which denoted by red color. It is grouped together which shows that the companies close to each other on the MDS plot are similar in the higher dimensional plane. Moreover, we can identify that cluster 3 is close to cluster 2.



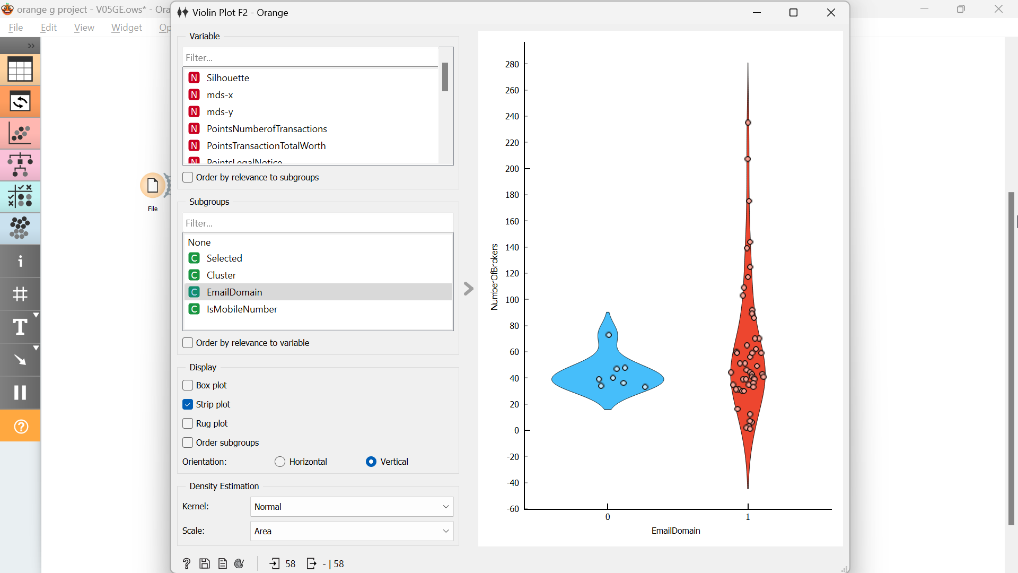
**Figure 6.** Strip plot visualization, where x axis denotes whether companies have email domain or not, and y axis denotes the DaysSinceLicense.

Figure 6 shows a strip plot visualization and comparison of companies based on whether or not they have an email domain. The blue strip is for companies which do not have an email domain and the red plot is for those with an email domain. The Y axis is showing DaysSinceLicense and each circle represents a company. The value range for the red strip plot (companies with email domains) is much larger than the value range for the other group. Furthermore, there seems to be many observations within the red strip that have higher values than the observations in the blue strip. Our observations from Figure 6 calls for the application of proper statistical test for the comparison of means of the two groups.



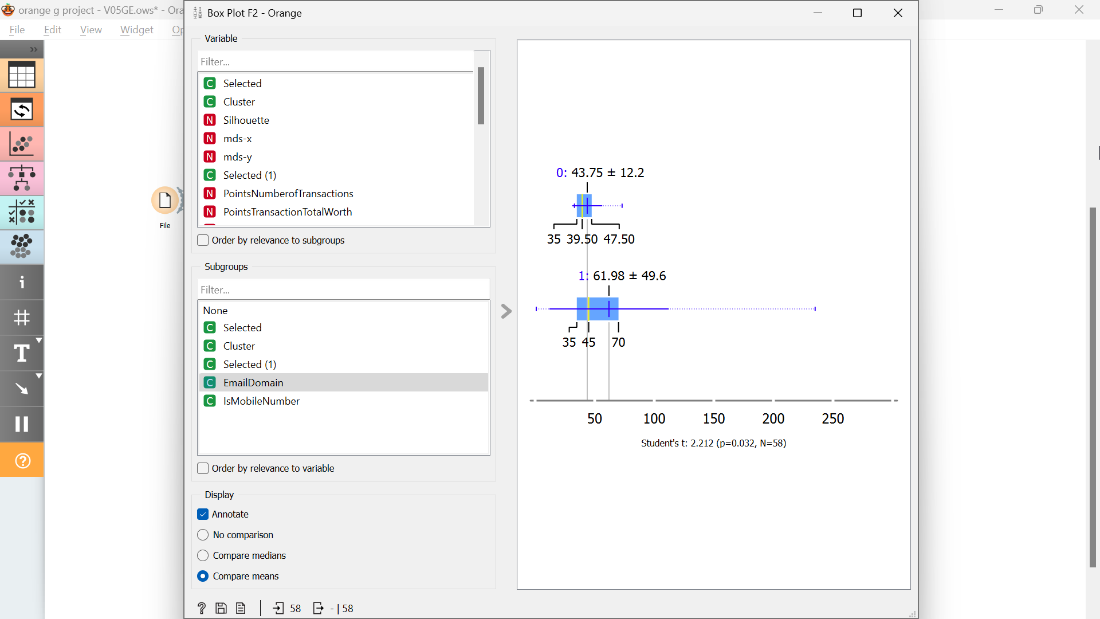
**Figure 7.** Box plot and statistical test for differences in means, as a complement to the analysis in Figure 6.

From the boxplot in Figure 7, we can also observe that the range and interval in the companies who have no email domain are narrower, while the range and interval on the companies who have an email domain is larger. Figure 7 also displays the results of student’s t test with a value of p= 0.001 obtained, suggesting strong statistical support for the differences in means. In other words, companies with an email domain name have higher values of DaySinceLicense.



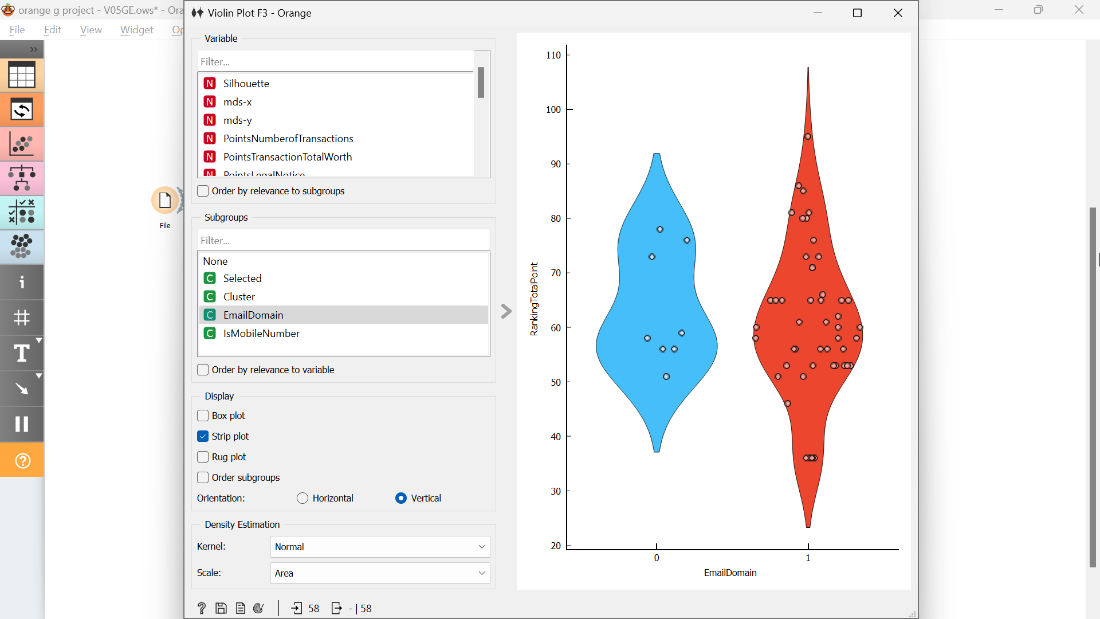
**Figure 8.** Strip plot visualization, where x axis denotes whether companies have email domain or not, and y axis denotes the NumberOfBrokers.

Figure 8 shows comparison of NumberOfBrokers for companies without and with email domains, shown with blue and red strip plots respectively. The red strip plot (companies with email domain) has larger range than the other, however, the higher values seem to be mostly outliers.



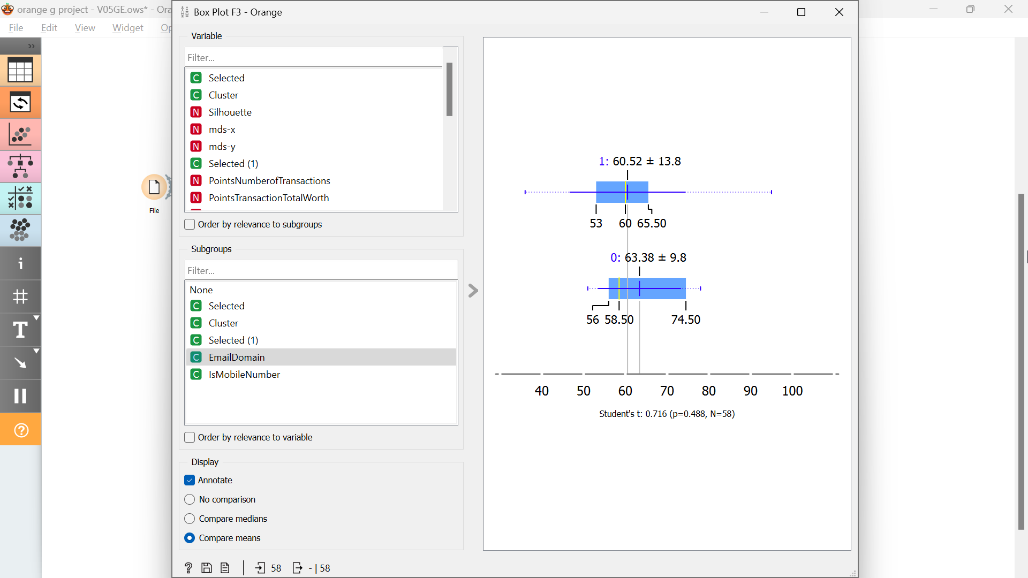
**Figure 9.** Box plot and statistical test for differences in means, as a complement to the analysis in Figure 8.

From the boxplot in Figure 9, we can see that the range of values for the companies who have an email domain is a bit higher than the one in the other group (ones who do not have an email domain) since they are narrower. Figure 9 also displays the student’s t test with p = 0.032 which suggests statistical support for the differences in means, however it is not as strong as the previous case.



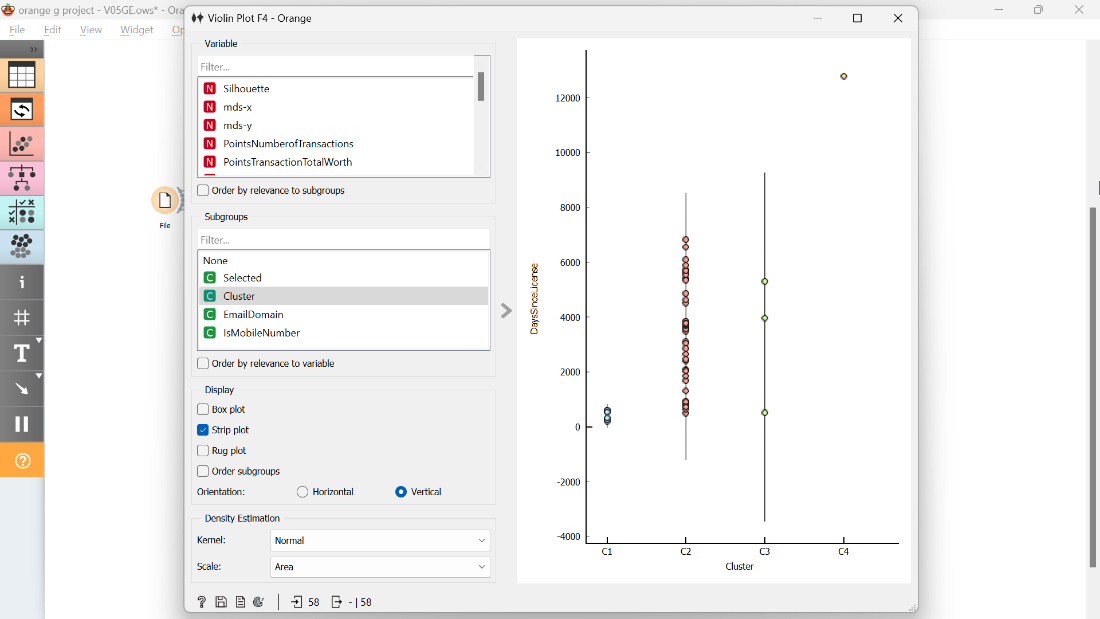
**Figure 10**

Figure 10 shows RankingTotalPoint on the Y axis while it shows the EmailDomain on the x axis. In Figure 10, the range of values for the red strip plot (companies with email domain) span beyond the blue strip plot (companies without an email domain). However, the central values for both plots seem almost same.



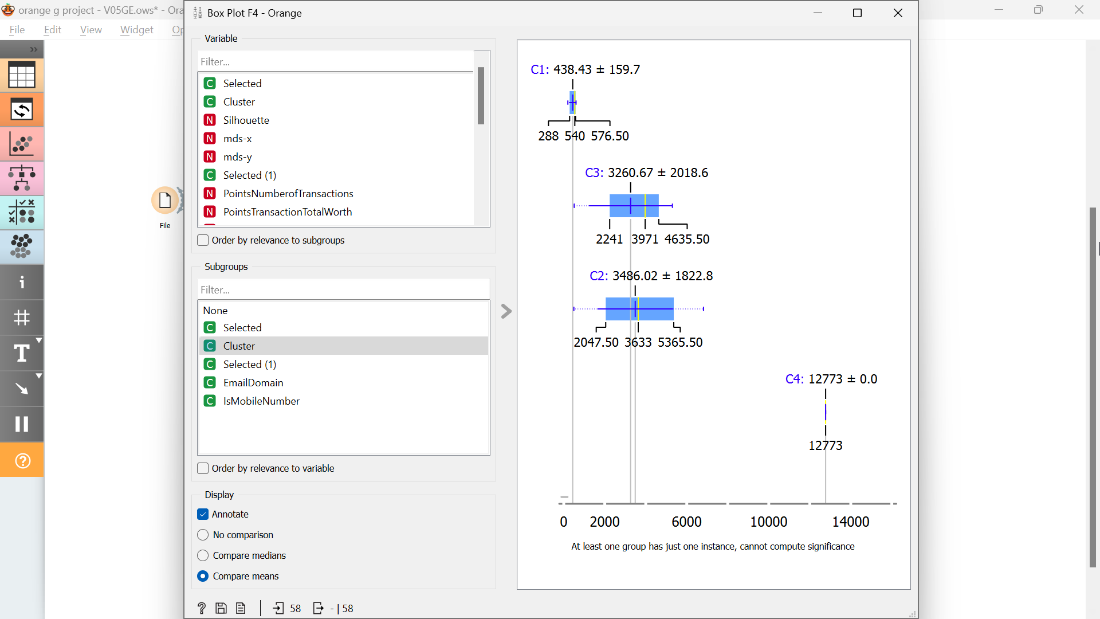
**Figure 11**

The boxplot in Figure 11 suggests the same as Figure 10. Furthermore, the application of student’s t test results in p= 0.488, which suggests that there is no significant difference in means.



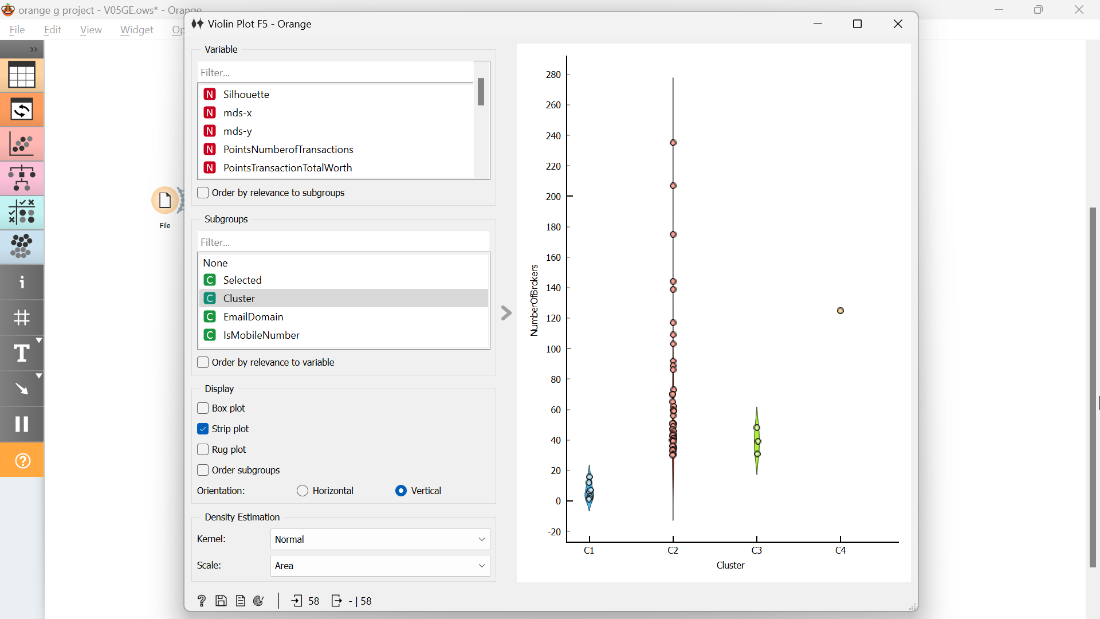
**Figure 12**

Figure 12 shows office DaysSincelicense on the Y axis while it shows the number of Cluster on the x axis. Cluster 4 shows an old company between the others, while cluster 2, and 3 shows younger companies than cluster 4. Moreover, cluster one shows new companies.



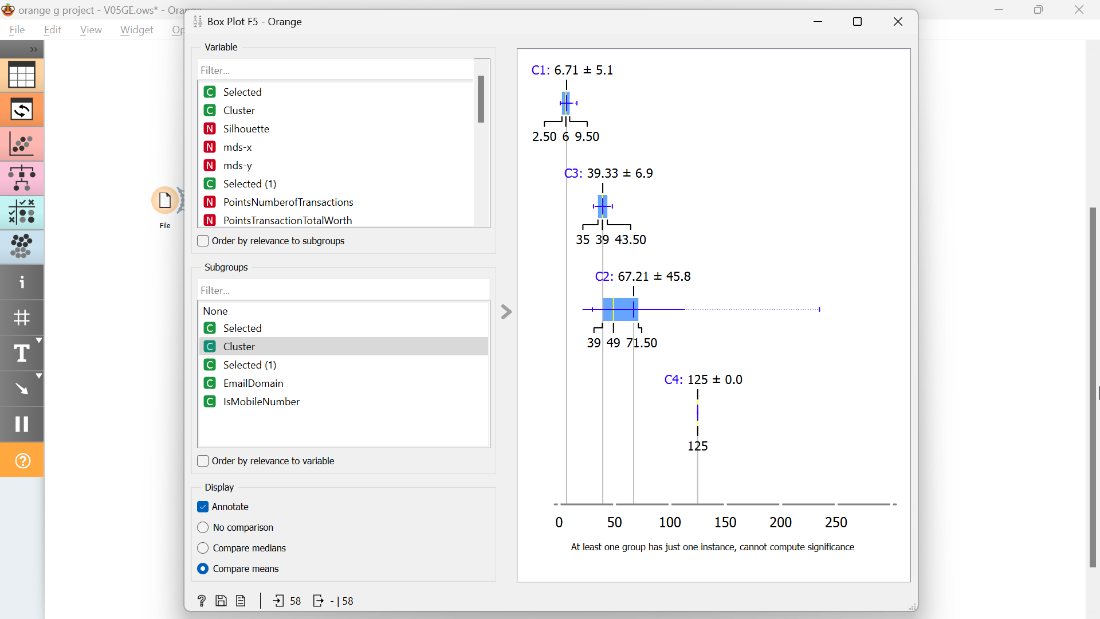
**Figure 13**

The boxplot in Figure 13 suggests the same as Figure 12. Furthermore, we can observe that cluster 2 has higher mean and narrower interval than cluster 3.



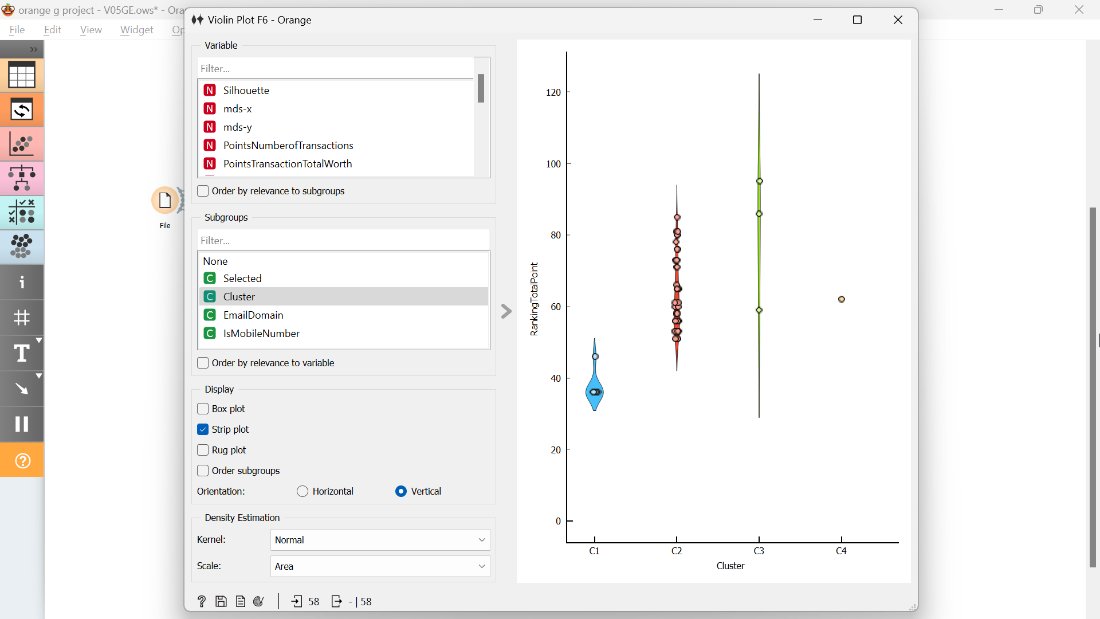
**Figure 14**

Figure 14 shows NumberOfBrokers on the Y axis while it shows the number of Cluster on the x axis. Stripplot shows that Cluster 2 has the highest range, while cluster 1 has the lowest.



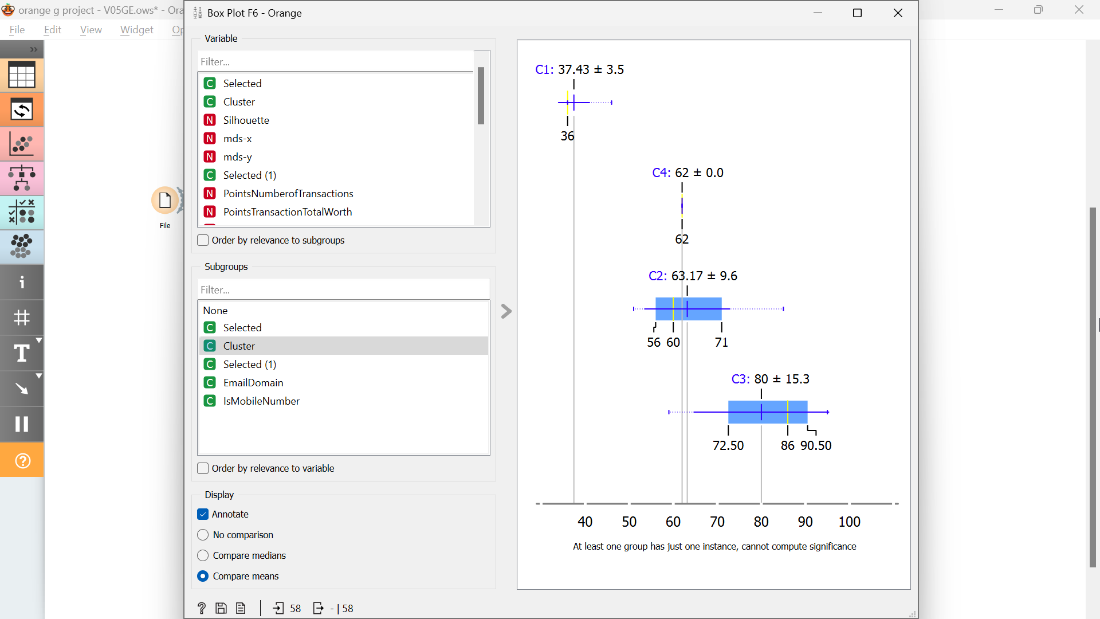
**Figure 15**

The boxplot in Figure 15 suggests the same as Figure 14. Furthermore, we can observe that cluster 4 has the highest mean and narrowest interval since it has single observation.



**Figure 16**

Figure 16 shows RankingTotalPoint on the Y axis while it shows the number of Cluster on the x axis. Stripplot shows that Cluster 3 has the highest range, while cluster 1 has the lowest.



**Figure 17**

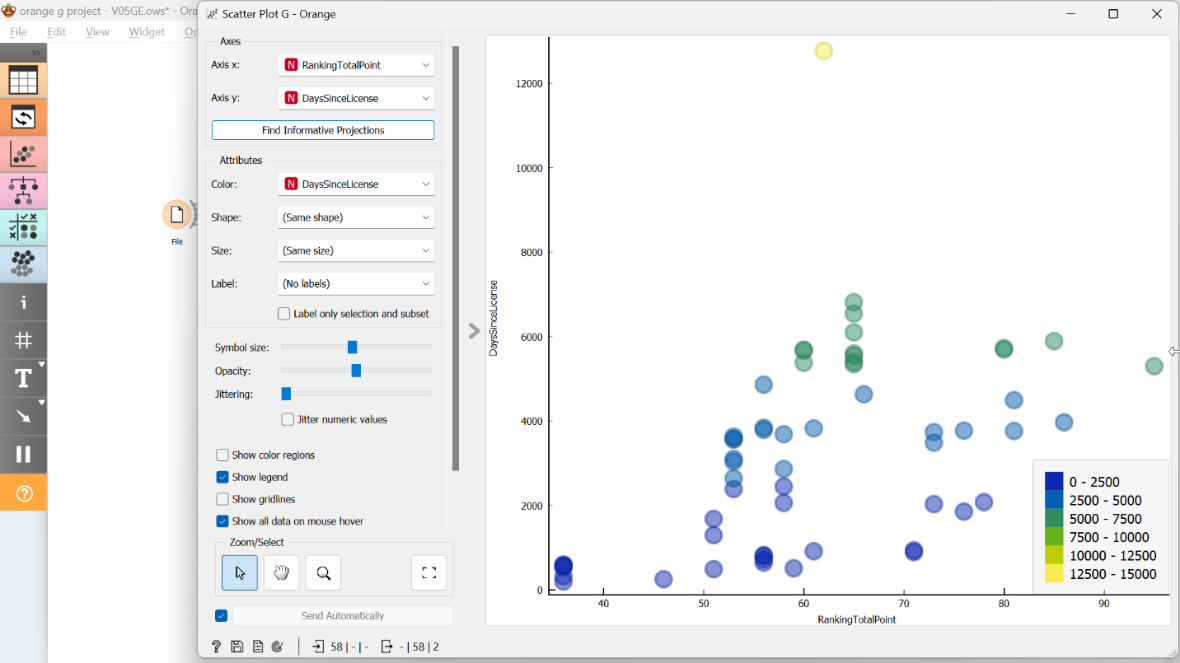
The boxplot in Figure 17 suggests the same as Figure 16. Furthermore, we can observe that cluster 3 has the highest mean and widest interval. Moreover, cluster 1 has the lowest mean. Also, cluster 2 and 4 have similar mean.

Graphical user interface

Description automatically generated

**Figure 18.** Columns selected for Scatterplot G, outlier analysis, and regression analysis

Figure 18 shows the columns selected for Scatterplot G, outlier analysis, and regression analysis.



**Figure 19.** Scatter plot of select attributes before regression.

* Figure 19 shows a scatterplot of the selected attributes. We can observe a relationship between RankingTotoalPoints and DaysSinceLicense being positive and somewhat linear. In Figure 19 we can observe that there is an outlier observation, represented by the yellow node on the top of the plot. Therefore, to be able to conduct a more reliable analysis, it is important to remove outliers.

Graphical user interface, text, application, email

Description automatically generated

**Figure 20.** Parameters for outlier analysis.

Outlier analysis is carried out using the Minkowski metric, with the other parameters also shown in Figure 20. There are five outliers identified, being BETTER, DACHA, HAMPTONS, HOMEMATTERS, and A1.

Graphical user interface, application

Description automatically generated

**Figure 21.** Regression coefficients.

Figure 21 displays the results of linear regression, conducted using “leave one out” training/testing split. R-squared value is 0.313: Indicating that the total variation of the data explained using regression analysis is 31.3%. While this is not nearly as high as one would hope to see, it can still be statistically significant. Further analysis using proper statistical tests is needed.

**Conclusions**

After applying various studies, we developed an analytics workflow with a clear understanding of interpreting brokerage companies in the real estate sector. To summarize our main breakthroughs, it was necessary to clean our dataset and identify some relationships between our dataset’s attributes. Moreover, many patterns and trends were spotted after using our statistical methods that would help any reader or a real estate buyer to make decisions regarding choosing a broker. It is important to highlight that we used Orange software features to help us analyze our data. After developing our first scatterplot, we concluded that a real estate buyer would favor brokers with high-ranking points and with an also high number of awards. Moreover, a hierarchical clustering was made so we selected k=4 as the number of clusters after going through our Silhouette scores. Furthermore, we used the multi-dimensional scaling (MDS) technique to represent complex, multi-dimensional data in a simplified two-dimensional Cartesian plane. this mapping method allows for a clearer visualization and easier interpretation of the data by reducing its dimensionality while preserving its inherent structure. In addition, we created more than one strip plot to show some comparisons that will help many real estate buyers to see the differences if different attributes were allocated. On top of that, we utilized the usage of boxplots to obtain different values of the p-value that will affect our statistical support. By using the Minkowski metric, We also used outlier analysis to see that we gained 5 outliers. It is worth noting that we used linear regression models and interpreted them by using test and score results.

The validity of our research could be affected by some threats, which can be addressed as the following:

* It is necessary to note that this dataset was analyzed in May 2023 since there would be future changes that could affect the credibility of our research. For instance, some offices might win awards in the following next years.
* For some broke offices, some missing values that were not found in the original dataset such as an office phone number.
* It is essential to acknowledge that the following study was based on Dubai real estate broker market whereas results may vary based on the city’s location.
* There might be a chance that there are some real estate brokers who might not be included in our dataset which could also compromise the reliability of the research.

**References**

1. Dubai Land Department. (May 10, 2023). Licensed Real Estate Brokers. Retrieved from <https://dubailand.gov.ae/en/eservices/licensed-real-estate-brokers/licensed-real-estate-brokers-list/#/>
2. Obinna, W. K., & Udo, M. J. (2022). Improving online Real Estate Management System using data analytics. *Journal of Emerging Technologies*, *2*(2), 66-75
3. Razen, A., Brunauer, W., Klein, N., Kneib, T., Lang, S., & Umlauf, N. (2014). *Statistical risk analysis for real estate collateral valuation using Bayesian distributional and quantile regression* (No. 2014-12). Working papers in Economics and Statistics.‏