Project - Descriptive Statistics

The following document will contain the Descriptive Statistics project of Profession AI's Master of Science in Data Science.

Task 1 - The dataset

The dataset concerns the sale of real estate in Texas.
The dataset consists of eight variables:

- city: city
- year: reference year
- month: reference month
- sales: total number of sales
- volume: total value of sales in millions of dollars
- median_price: median selling price in dollars
- listings: total number of active listings
- months_inventory: amount of time needed to sell all current listings at the current sales rate, expressed in months

The import of data into R was carried out using the following line of code:
data_texas <- read.csv("realestate_texas.csv")
which made it possible to read the entire dataset and place it within the R environment.

Task 2 - Description of the dataset

The dataset is explored using R's basic 'head' function, which allows us to display the rows and columns of the dataset. The line of code used was:

head(data_texas, 5)

Below is an analysis of each variable in the dataset:

- city: qualitative variable on a nominal scale
- year: qualitative ordered variable
- month: ordered qualitative variable
- sales: discrete quantitative variable
- volume: continuous quantitative variable
- median_price: continuous quantitative variable
- listings: discrete quantitative variable
- months_inventory: continuous quantitative variable

All in all, there are 2 discrete quantitative variables, 3 continuous quantitative variables, 1 qualitative variable on a nominal scale and 2 qualitative ordered variable.

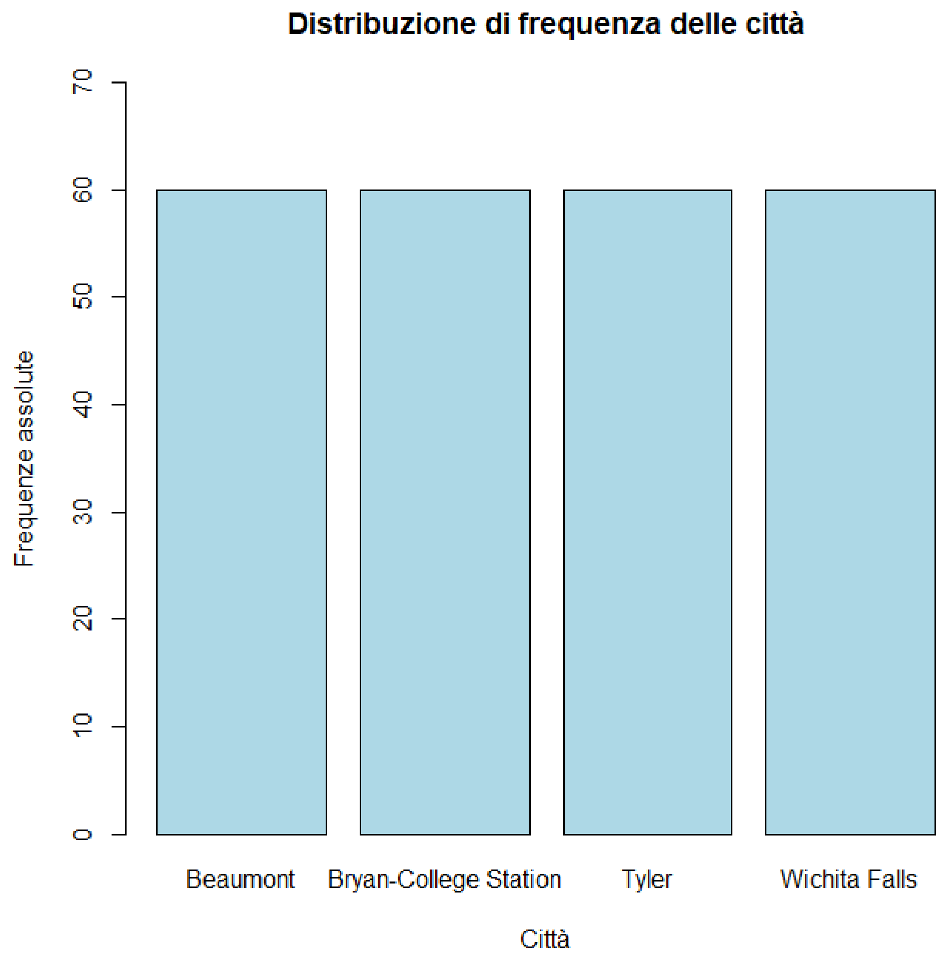
Task 3 - Calculation of indices for variables

The variables identified for which it does not make sense to calculate the position, variability and shape indices are:

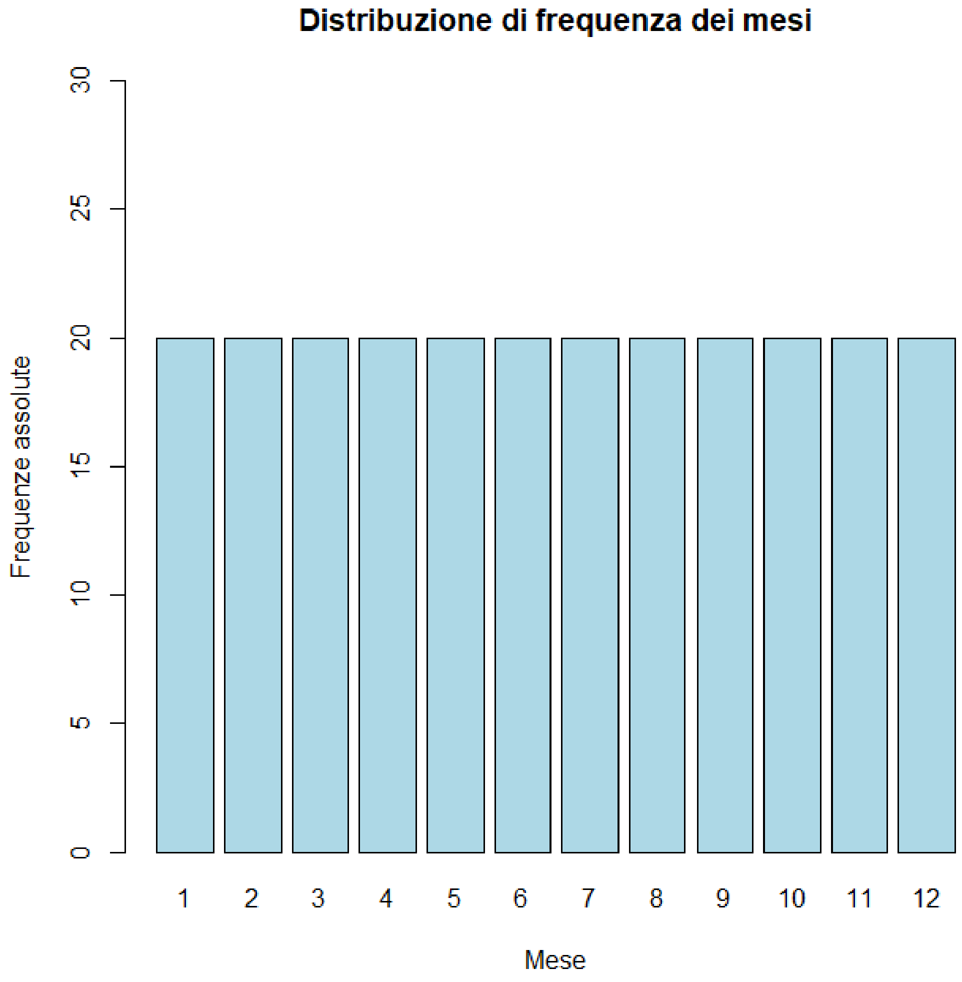
- city: qualitative variable on a nominal scale
- year: ordered qualitative variable
- month: ordered qualitative variable

This is because they do not give additional information by studying the indices as variables that already have a natural order of their own. Moreover, going to calculate an average, for example, would not be correct because although years and months are numeric variables, they cannot be summed up and thus, summarised. For these variables, frequency distributions were calculated and visualised using a barplot-type graph from the basic R package.

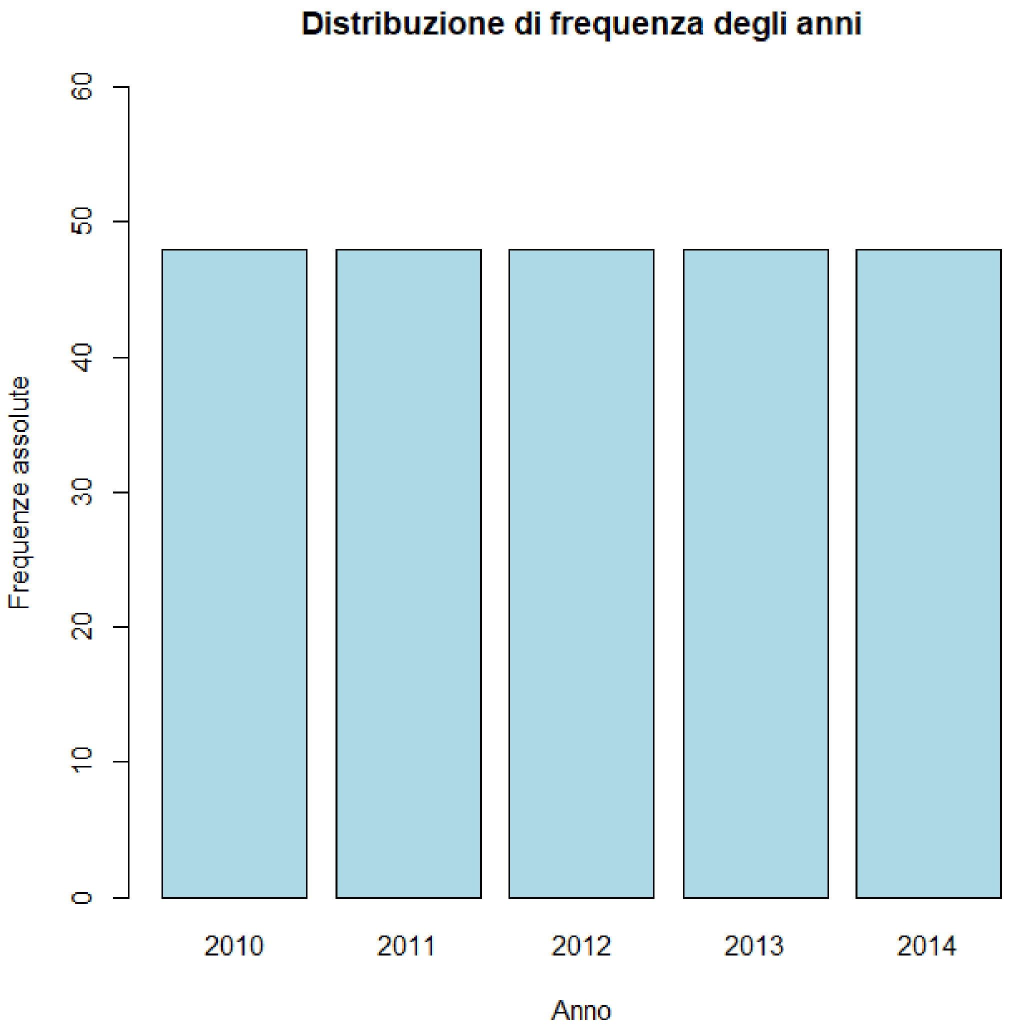
Frequency distribution for the variable city



Frequency distribution for the variable month



Frequency distribution for the variable year



The variables useful for calculating the position, variability and shape indices are:

- sales: total number of sales
- volume: total value of sales in millions of dollars
- median_price: median selling price in dollars
- listings: total number of active listings
- months_inventory: amount of time needed to sell all current listings at the current sales rate, expressed in months

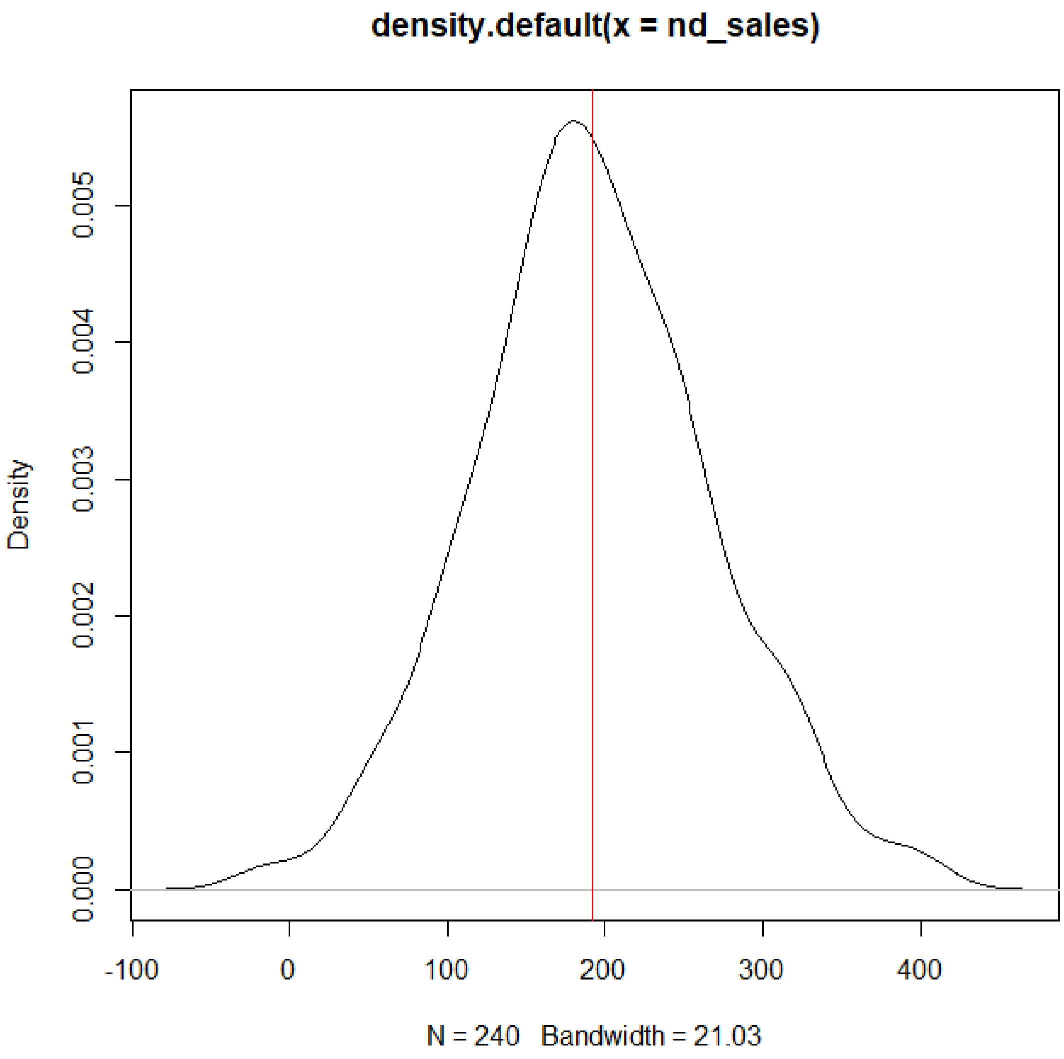
Of these, position indices were calculated individually through R basis functions, variability indices through custom implemented functions and finally, shape indices by constructing the normal distribution of the variables.

In addition, the normal distribution was plotted as a probability density graph.

Summary table of position, variability and shape indices for the variables: sales, volume, median_price, listings and month_inventory

	sales	volume	median_price	listings	month_inventory
Min	79	8.166	73800	743	3.4
Max	423	83.547	180000	3296	14.9
Mean	192.3	31.00519	132665.4	1738.021	9.1925
Median	175.5	27.062	134500	1618.5	8.95
Range	344	75.381	106200	2553	11.5
Variance	6317.865	276.1154	511433095.65972	564208.3	5.28477
Standard deviation	79.485	16.61672	22614.88659	43	2.29886
Coefficient variance	41.42203	53.70536	17.08218	43.30833	25.06031
Fisher index	0.71362	0.87921	0.87921	0.64544	0.04071
Kurtosis index	-0.33552	0.15056	0.64272	-0.81015	-0.19794

Density graph of the sales variable



Task 4 - Check which variable is found to have greater variability

The analysis on testing the variable with the greatest variability was to create boxplots of the individual variables considered assessable from the point of view of calculating indices, in order to compare the distributions. Subsequently, the standard deviation of each variable was compared and it emerged that in particular, 'listings' and 'median_price' had a standard deviation in orders of magnitude larger than the variables 'months_inventory', 'sales' and 'volume'. For this reason, it was deemed necessary to compare the coefficients of variation, which allow the variability to be assessed relatively between different distributions. In this way, it was possible to compare the variability and it emerged that the variable with the greatest variability is 'volume', which has a coefficient of variation of 53.7, i.e. a standard deviation of 57% of the mean.

As for the most asymmetrical variable, a comparison of the Fisher indices between the various variables was carried out and the one with a higher value was identified than all the others.

The comparison showed that the variable with the greatest asymmetry is: 'volume', which has an index value of 0.88. This index value indicates that the distribution of the variable 'volume' is positively skewed. In addition, when the Kurtosis index is also checked, the distribution also turns out to be leptokurtic as it has an index of 0.15 (thin tails and in the central area sharp, high curves).

Task 5 - Quantitative variable divided into classes

The variable chosen for the division into classes was 'sales'.

The choice of the classes of 0 - 500 with a step of 100 is because given the minimum of the distribution of 79 and the maximum of 429, it would have made no sense to create classes 0 - 500 with a step of 50, you would have obtained the first class 0 - 50 completely empty, which is why we opted for 0 - 500 with a step of 100.

A new column was added to the main dataset to include all classes of the sales variable so that its frequencies could be calculated later.

Once the classes were established and added to the dataset, the absolute, relative and cumulative relative frequencies were calculated for the classes constructed.
All frequencies were entered into a dataframe with which bar graphs displaying both absolute and relative frequencies divided by classes could be created

Finally, the Gini index was calculated for the variable Sales, which was found to be 0.9.
Its value suggests that the distribution of the continuous variable Sales appears to be almost equidistributed along its distribution.

Task 6 - Gini index for the city variable

Considering that by definition the Gini heterogeneity index measures the propensity of a qualitative variable to assume its different modes, for the city variable there is an equidistribution, i.e. the Gini index is exactly equal to 1.

This deduction arises when considering the absolute frequency of the city variable, which is 60 for each mode (for each city).

To perform this check, the 'table' function was used to display the absolute frequencies and then, as a further check, the calculation of the Gini index.

Task 7 - Calculation of probabilities

The probability calculation for the variables city, month and month versus year was to plot the relative frequencies for each variable. Being discrete variables, we proceeded with bar graphs in order to display their relative frequencies. The outcome of the plot of both variables was that: the variable city presents an evenly distributed relative frequency for each mode, since there are 4 cities in the dataset, the percentage with which the city "Beaumont" can be found from a random extraction is 25%, while for the variable month, the distribution behaves in the same way, but having 12 modes, the probability of extracting the month of July is 8% (see graphs below).

Finally, the same approach was used to calculate the probability of extracting the month of December 2012. The relative frequencies of the variables month and year together were calculated and then plotted in the form of both boxplots and barplots in order to assess the distribution of the two variables.

The outcome of the month variable was that it was evenly distributed along the years of the dataset, and that the probability of December 2012 coming out of a random draw is approximately 1.7%.

Task 8 - Average price column

The rationale for adding the new column plotting the average price for each month of the dataset was to consider the 'sales' and 'volume' columns corresponding to the total number of sales and the total value of sales in millions of dollars.

By relating the volume, i.e. the total value of sales expressed in millions of dollars, to the total number of sales, it is possible to obtain the average price for each month. In particular, since volume is expressed in millions of dollars, it was deemed appropriate to multiply it by one million in order to calculate the average price in dollars.

Task 9 - Sales Effectiveness

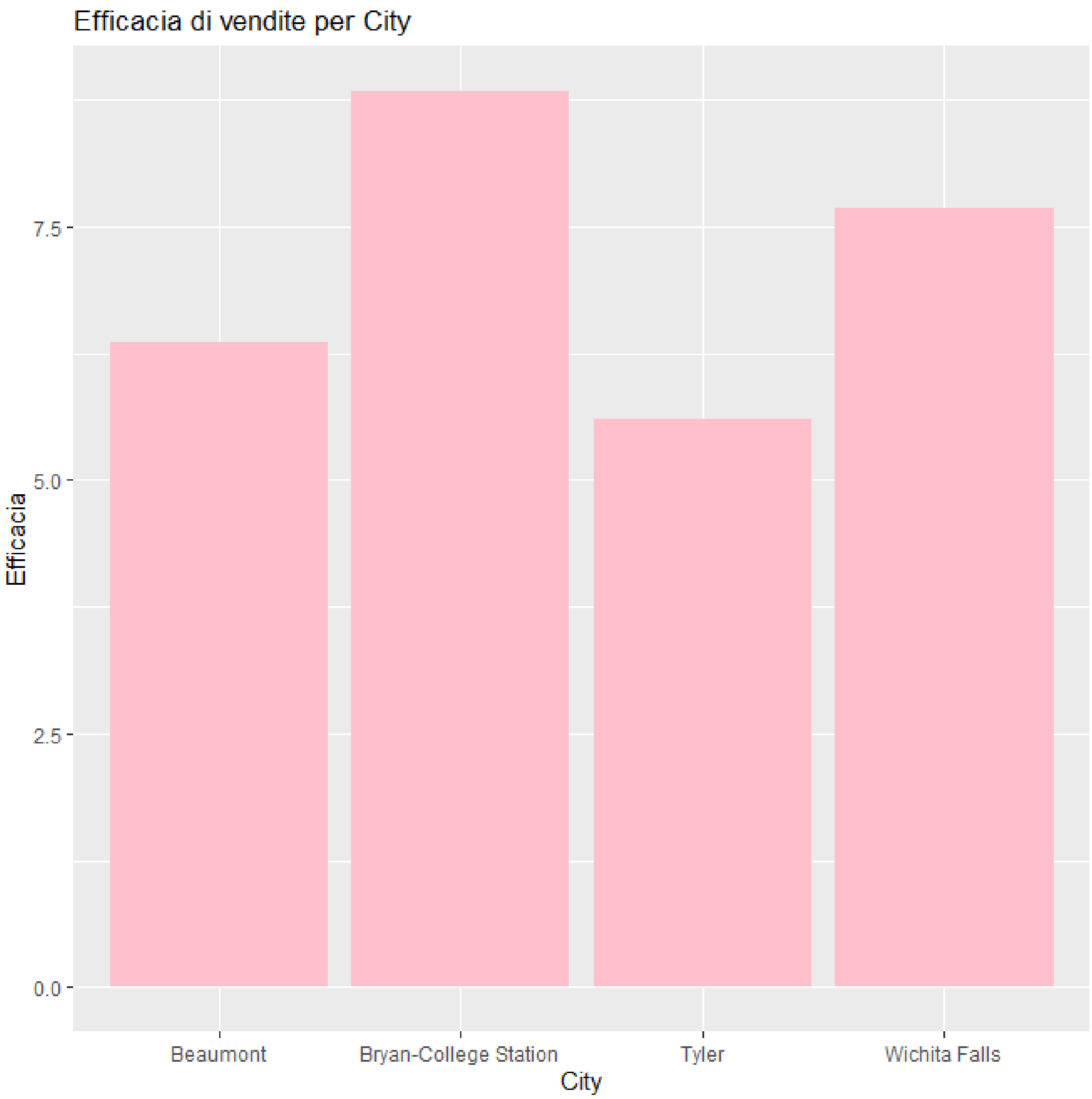
The effectiveness of sales was calculated on the basis of knowing the sales of real estate made and the properties still unsold. The ratio of sales made to properties still unsold gives information on how effective sales were in each city for each month of the years 2012 to 2014.

Once the effectiveness had been calculated, as described above, we also proceeded to assess where effectiveness was found most in relation to cities and years by plotting effectiveness against the city and year variables. It was considered appropriate to create an aggregation of effectiveness by city, month and year, and then go on to construct a geom_bar graph broken down by city, thus assessing sales effectiveness separately for each city, month and year.

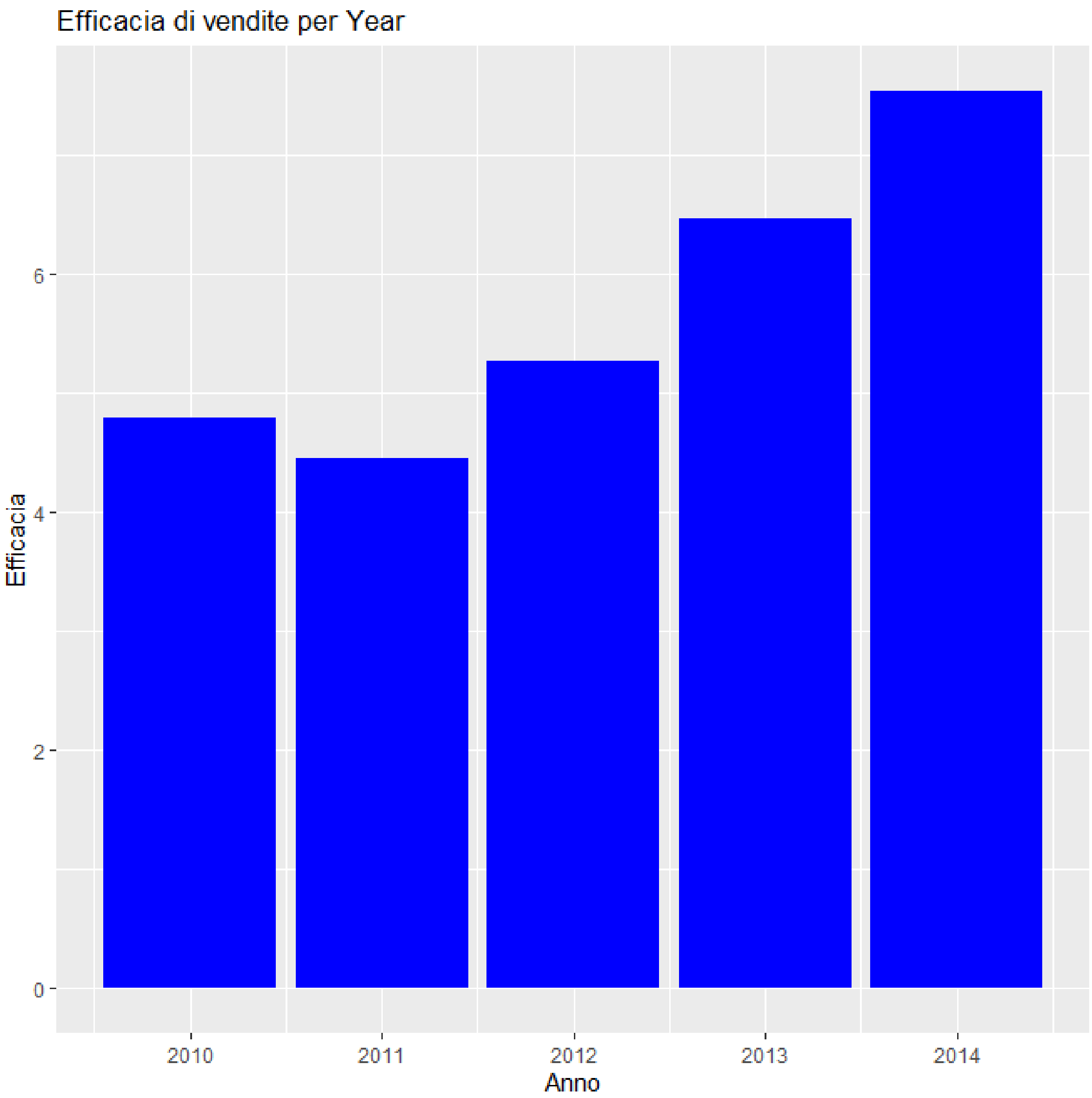
Subsequently, a 3-dimensional graph was created in order to report in which month of the year and in which city a higher percentage of sales was found and therefore a higher sales effectiveness.

The outcomes were that the greatest sales effectiveness was found in the city of "Bryan-College Station" and in the year 2014. In particular, sales were most successful in July of the year 2014 for the city "Bryan-College Station".

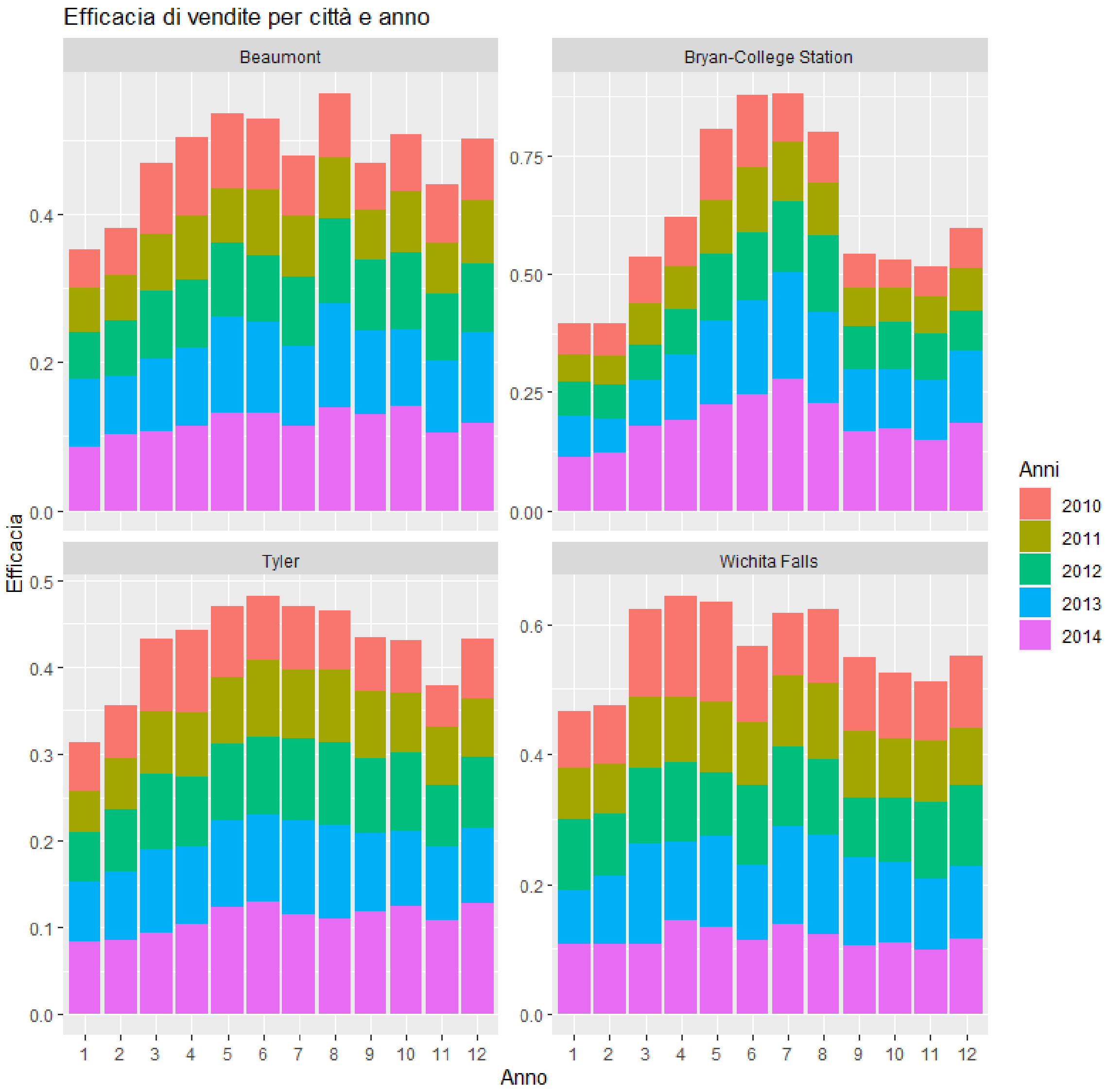
Sales effectiveness for City



Sales effectiveness per year



Sales effectiveness for Year, City and Month



Task 10 - Summary of sales, volume and listings variables

The variables chosen for the creation of the summarises using the 'dplyr' library were: sales, volume and listings. The objective was to evaluate their mean and standard deviation in order to draw conclusions on what was the trend in sales, dollars and listings still active.

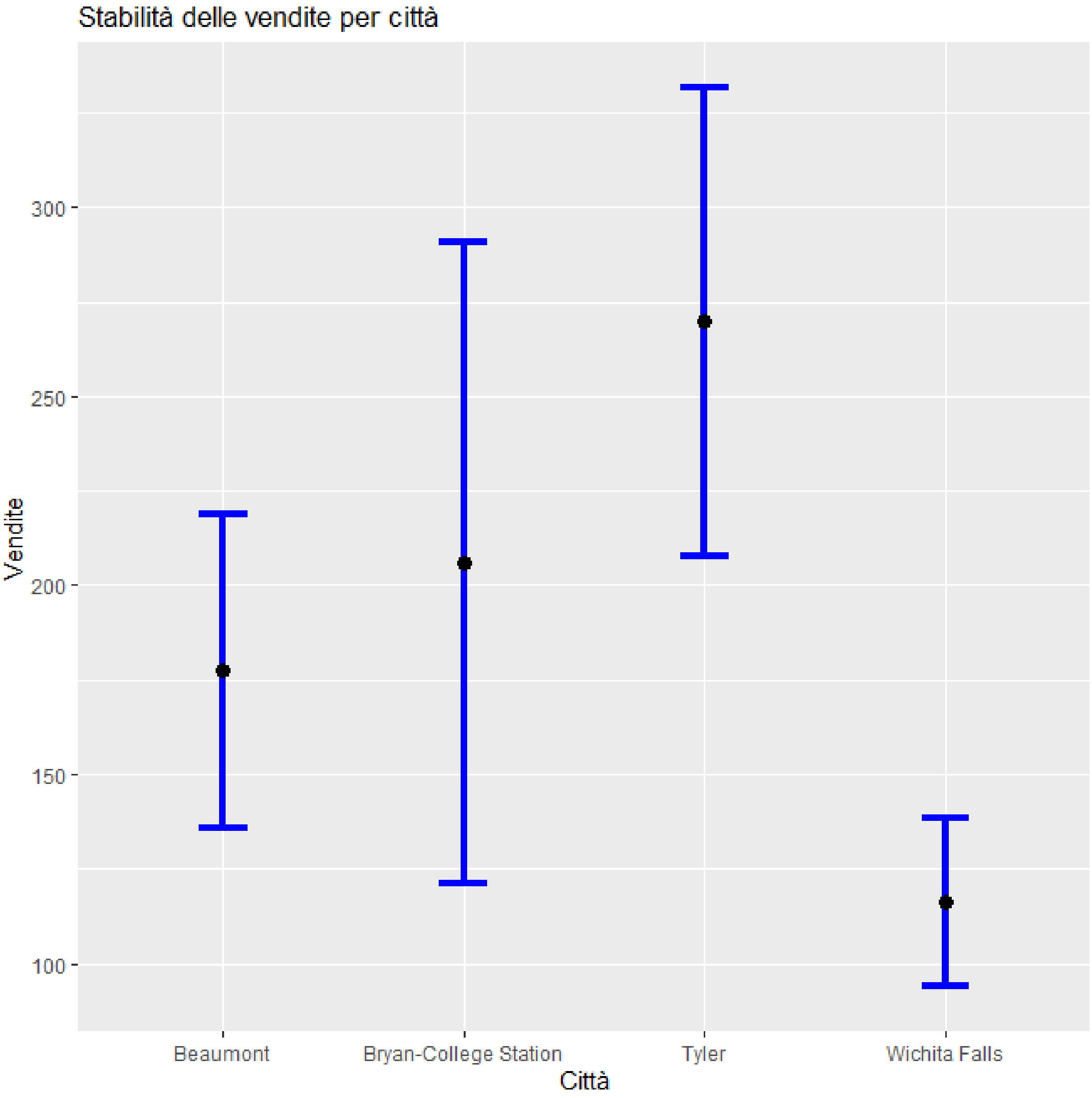
Therefore, graphs were created for some of the summarises created to draw the conclusions described in the objective.

The first bar graph (geom_errorbar) in which the mean and standard deviation of sales for each city was compared was done in order to understand in which city on average properties were sold the most and by how much they varied. The city of "Bryan-College Station" was found to be the city with an average of approximately 200 sales and with a very high variability, so this city shows little stability in sales compared to the average, in contrast to the city of "Wichita Falls" which shows greater stability with a much lower average of sales.

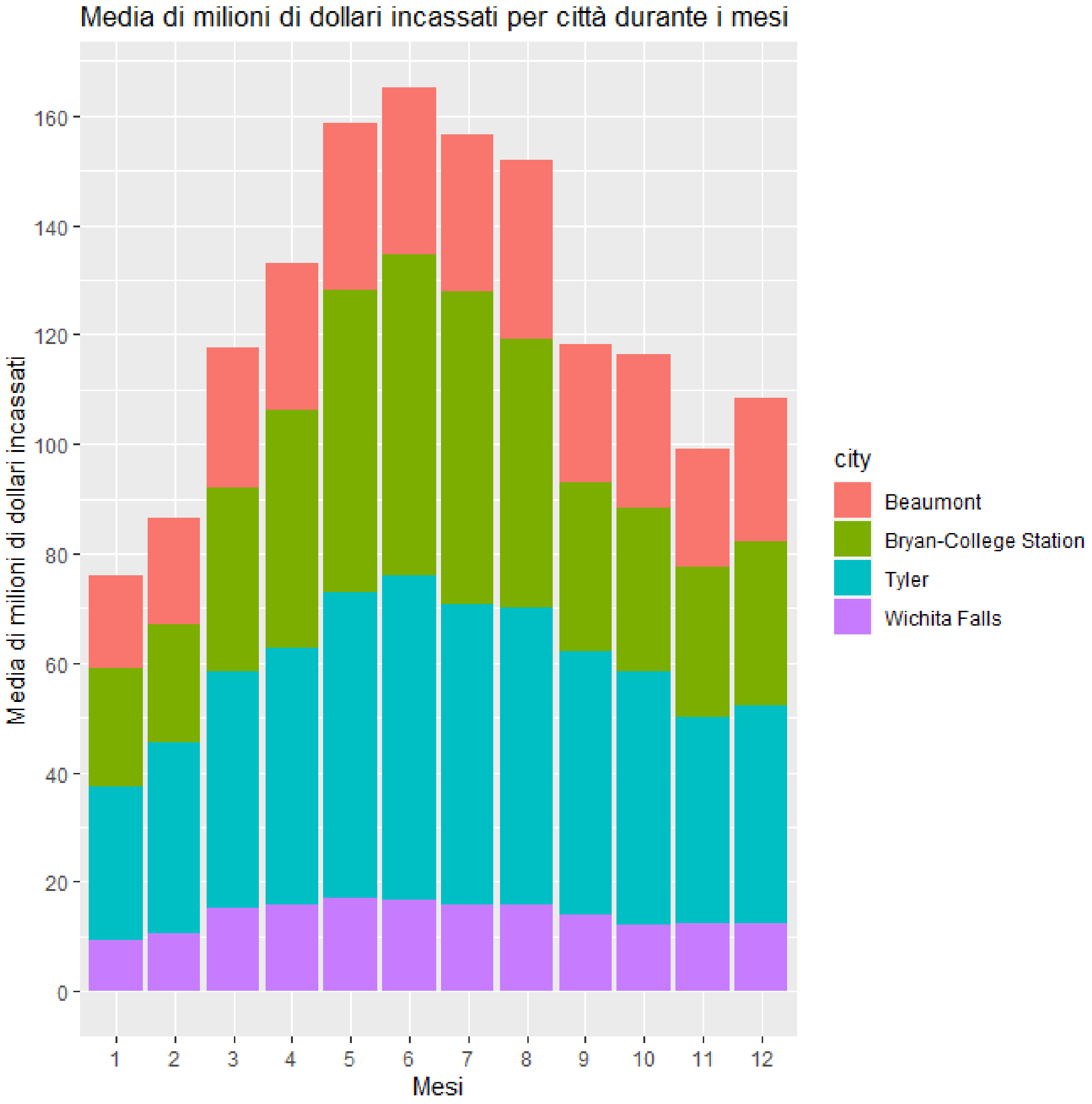
Next, it was decided to create a bar graph to assess in which months of the year, on average, real estate produces a higher profit and in which city. For this, the average millions of dollars were related to the months of the years and the corresponding cities. The result was that in the month of 'June', the average number of dollars collected from sales was higher than in all other months, and in particular, the highest receipts were found in the cities of 'Tyler' and 'Bryan-College Station'.

Finally, it was decided to create an overlapping bar graph to assess in which months of the years 2010 and 2014, reported from the dataset, sales occurred most on average. For this, the average sales were related to the years and months. The result was that from the year 2011 to the year 2014, the sale of real estate increased and in particular, in the year 2014 the most real estate was sold on average.

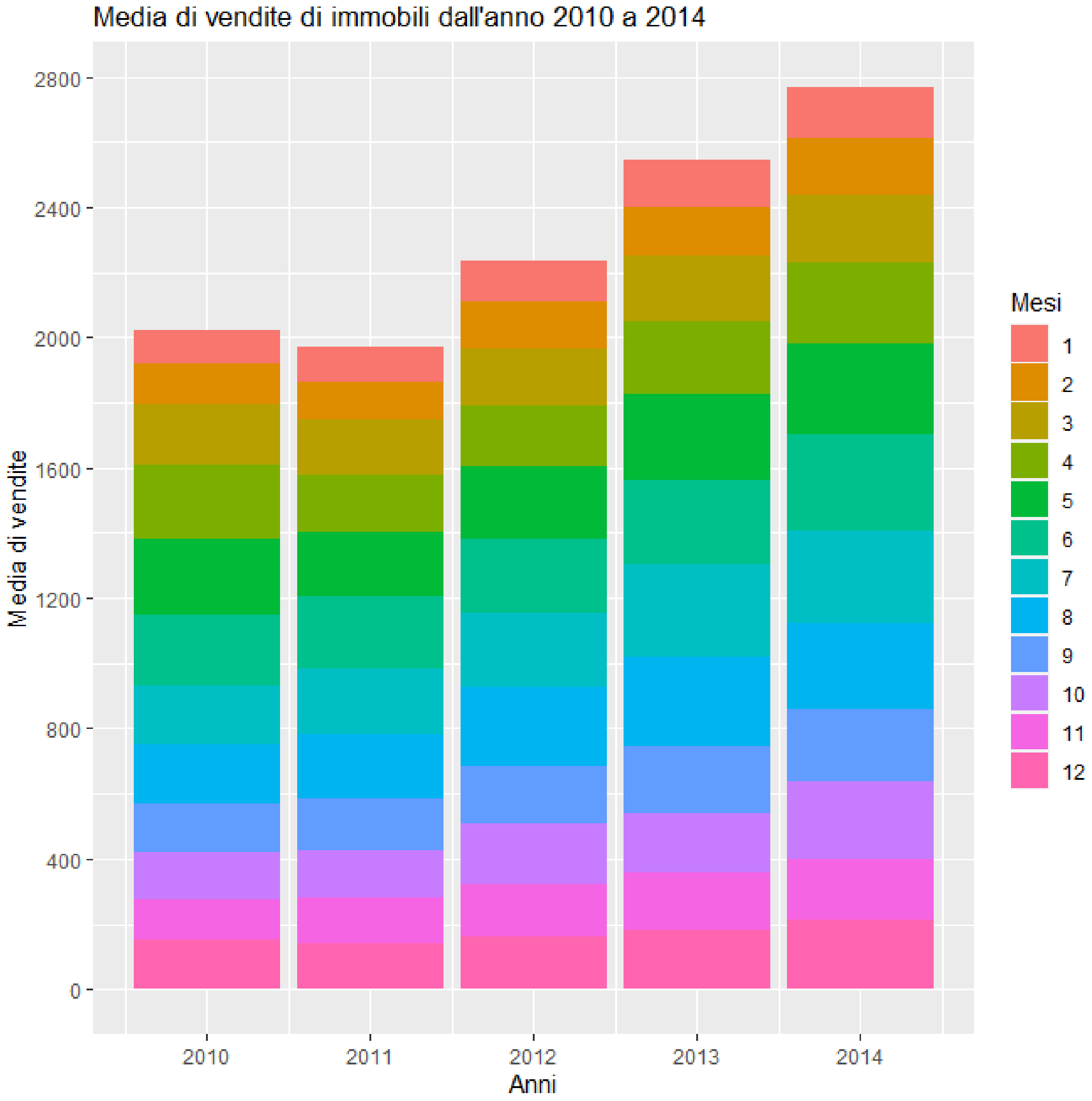
Stability of sales by city



Average millions of dollars collected per city over the years



Average of property sales from the year 2010 to 2014



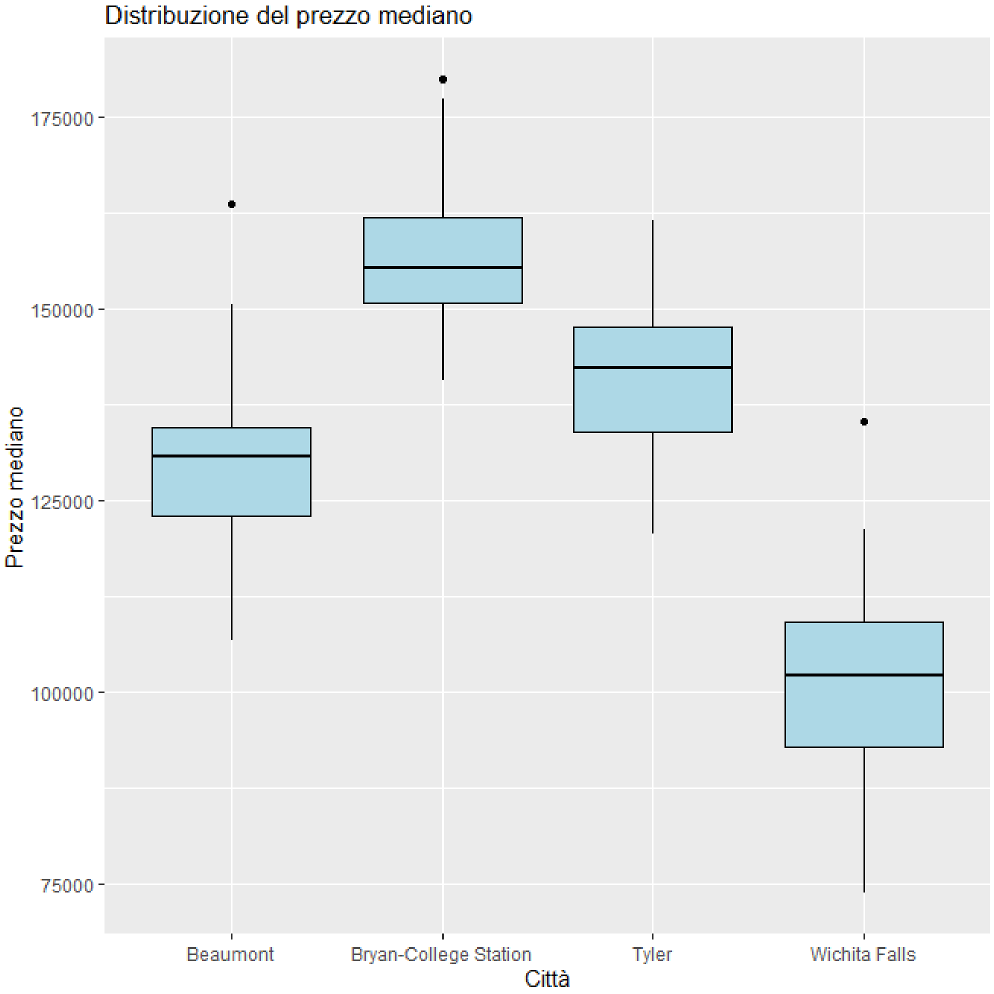
Task 11 - Boxplot median_price between cities

The comparison of the median_price between cities and between cities in relation to years was carried out by means of boxplot graphs summarising certain indices of position and variability with which to draw conclusions on the distribution of the median_price variable.

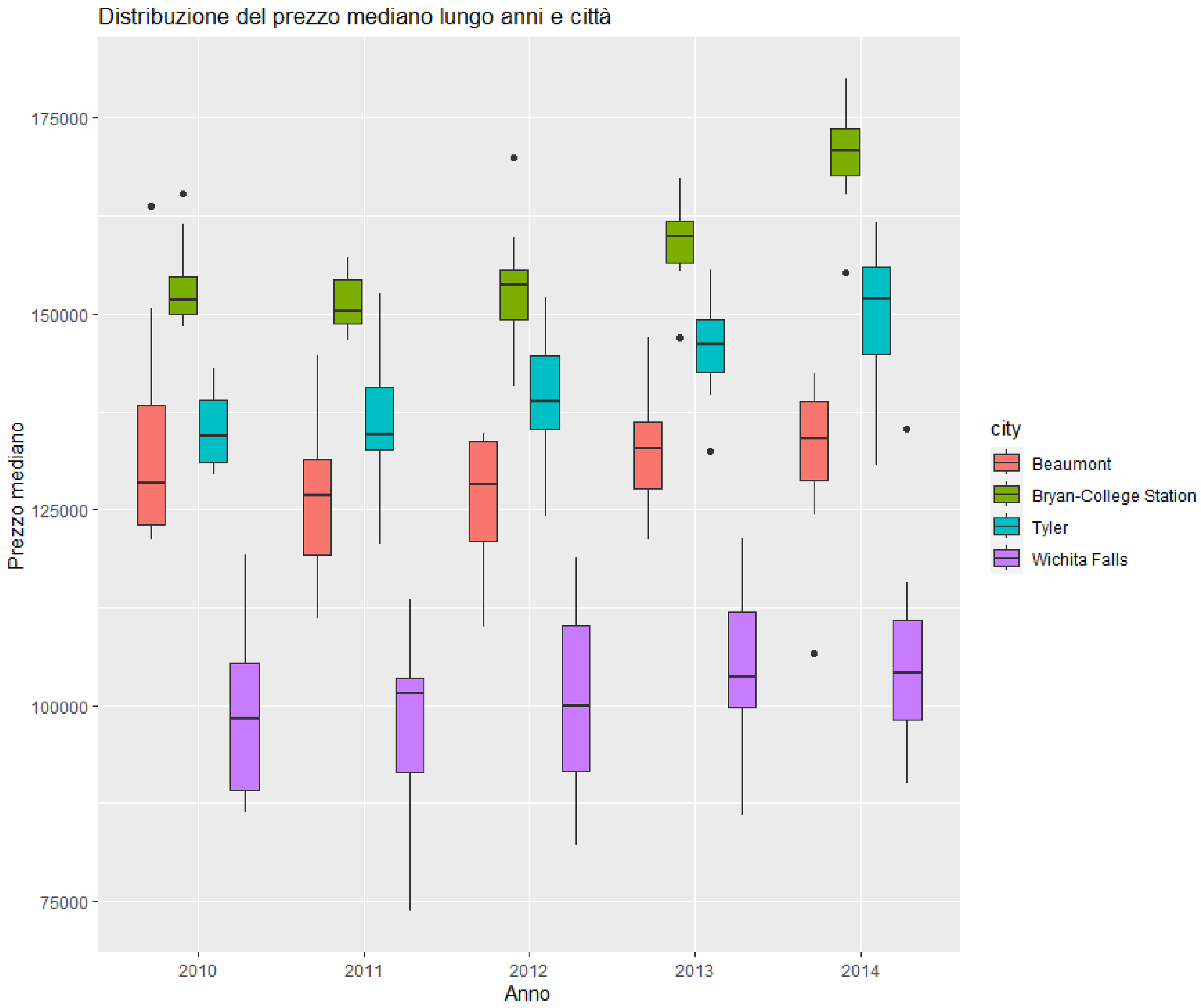
From the comparison of the boxplots of the median price per individual city, it can be seen for the city "Wichita Falls" that the variability of the median price is high compared to the other cities despite the fact that the median price has a much smaller range than the other cities. In the opposite case, on the other hand, it can be seen that the city "Bryan-College Station" presents a lower variability than the city "Wichita Falls" and its interquartile range is smaller, denoting a greater stability of the median price along the city in question.

From the comparison of the boxplots of the median price per city conditional on the years, considered as factors, it can be deduced that during the years from 2010 and 2014, for the city "Wichita Falls" the median price varied a lot presenting even an outline value in the year 2014. This may mean that in the city of "Wichita Falls" the median price is not very stable and varies a lot compared to other cities. In the opposite case, on the other hand, it can be seen in the city of "Bryan-College Station" how the median price, from 2010 to 2014, is more or less constant and does not show a high variability. In addition, it can be seen that in 2014, the median price increased by quite a bit compared to previous years.

Median price distribution along cities



Comparison of median price variability across years and cities

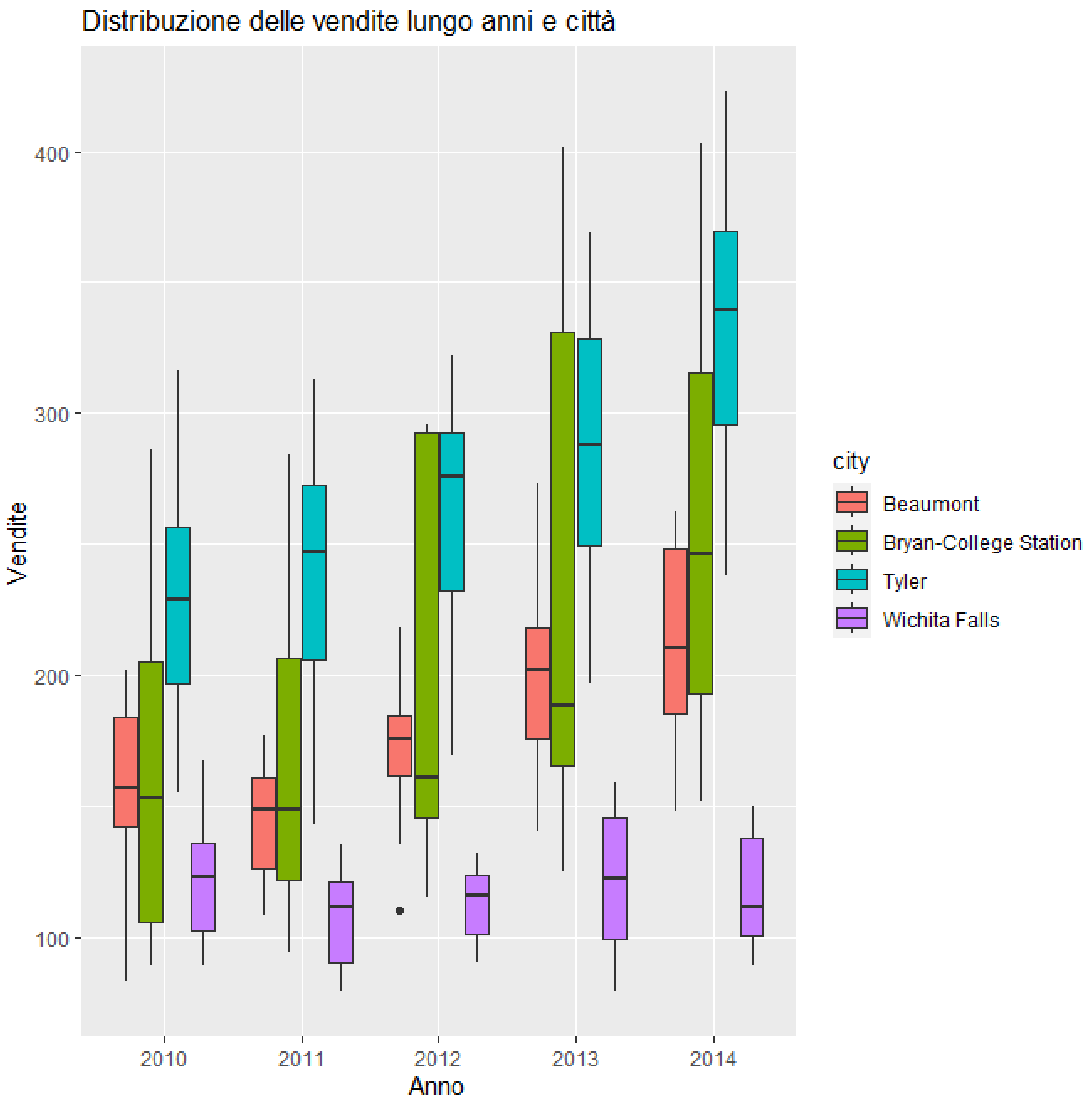


Task 12 - Boxplot sales between cities

The evaluation of sales during the years 2010 to 2014 was carried out by comparing the boxplots of sales conditional on each individual city for each year from 2010 to 2014. Subsequently, an overlapping bar graph was created with the same relationship between the variables sales, city and year.

From the boxplots, it could be seen that the variability of sales in the city "Bryan-College Station" over the years has increased over time and, in particular, in the year 2013, they fluctuated greatly compared to the years before and after. The city that has been most successful in sales is "Tyler", which over the years has shown continuous growth in its range, even reaching a sales quota of over 400. Finally, the city that has remained more or less constant over time, but which showed a decline in sales in 2014 was 'Wichita Falls'.

Comparison of median price variability across years and cities



Task 13 - Overlapping bar graph for sales over years and months

The superimposed bar graph for the evaluation of property sales conditional on years, months and cities was created by dividing a graph for each city in which the number of sales for each year was plotted for each month with a blue colour scale to highlight and mark the different months.

The graph shows for each city:

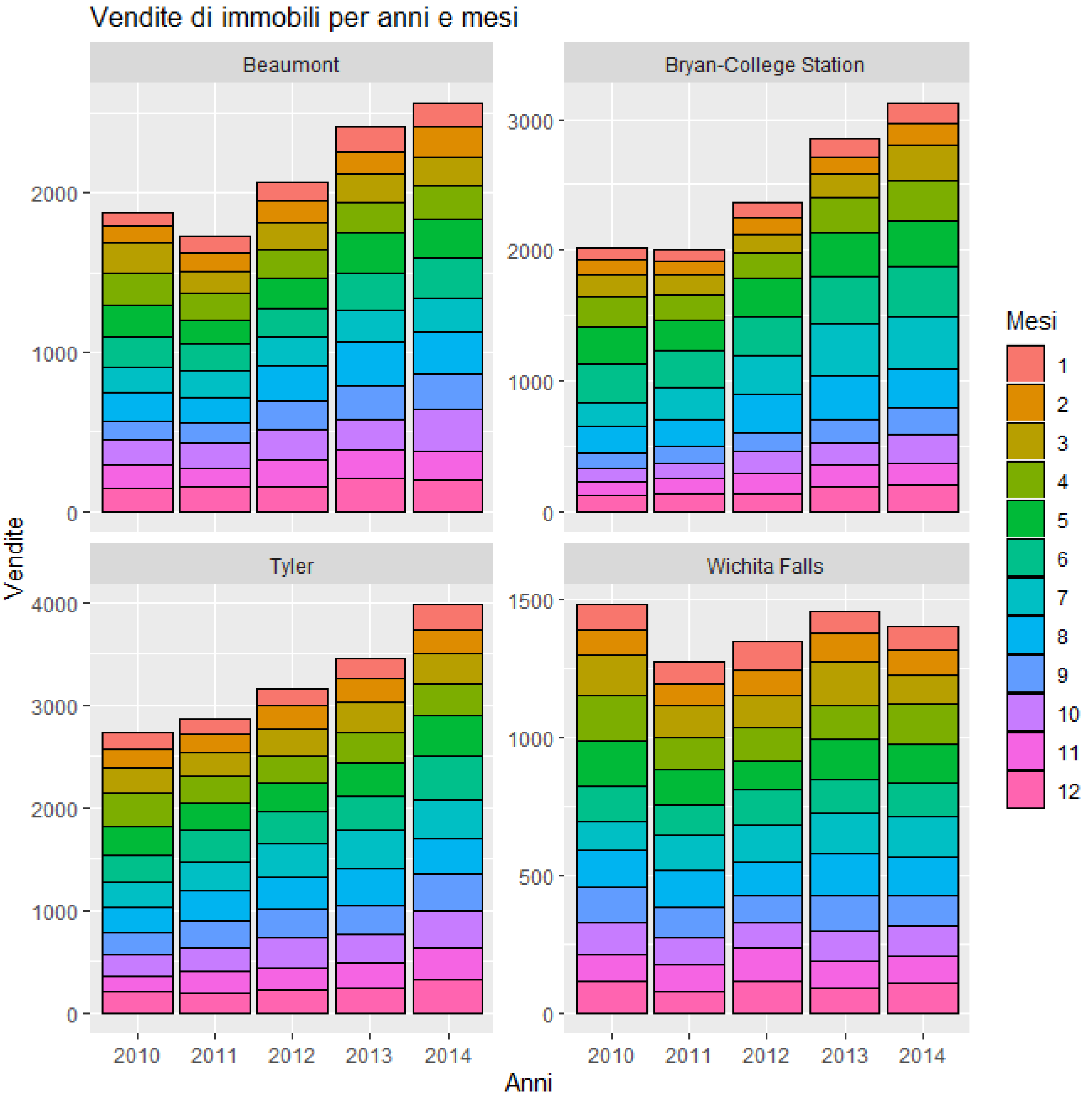
- **Beaumont**: sales dropped in 2011 compared to all others and in 2014 there was an increase in sales
- **Bryan-College Station**: sales have been more or less constant and increasing over the years, peaking in 2014. In particular, the highest sales were recorded during June and July.
- **Tyler**: sales have been growing steadily and in fact is the city where most properties have been sold. In particular, it can be seen that sales were also constant during all months for each year respectively

- **Wichita Falls**: sales have fluctuated quite a bit over the years, peaking in 2010 and then declining in 2011. In particular, it can be seen that sales were highest during the months of May and July.

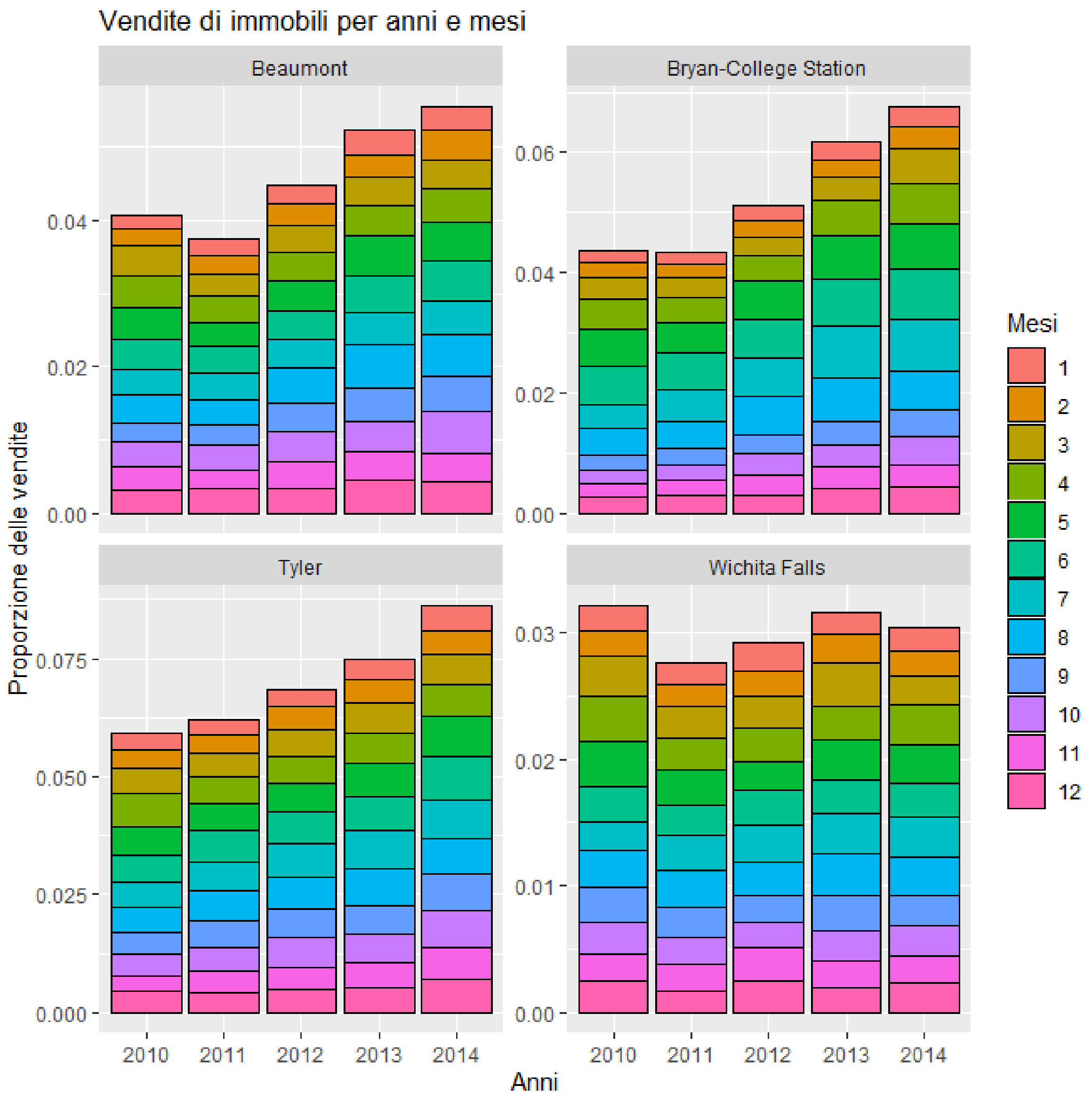
From the normalised bar graph, it is possible to draw conclusions on sales conditional on each month and year, compared to total sales in all cities, in all years and months.

Confirming the evidence from the non-normalised bar graph, it can be seen that for the cities of "Bryan-College Station" and "Tyler", sales accounted for approximately 6% and 8%, respectively, of total sales in the year 2014, thus representing the cities with the highest number of sales in that year; while for the cities of "Beaumont" and "Wichita Falls", sales during the years 2010 to 2014, accounted for approximately 3% and 5%, respectively.

Property sales over years and months in cities



Property sales over years and months in cities with normalised bar graph



Task 14 - Line chart

The realisation of the line chart was performed on the variable "volume" of the dataset, which represents in millions of dollars, the money collected from the sales of real estate. The idea was to plot along each year, the volume of dollars collected for each month and highlight the various cities through colours.

The graph shows for each city

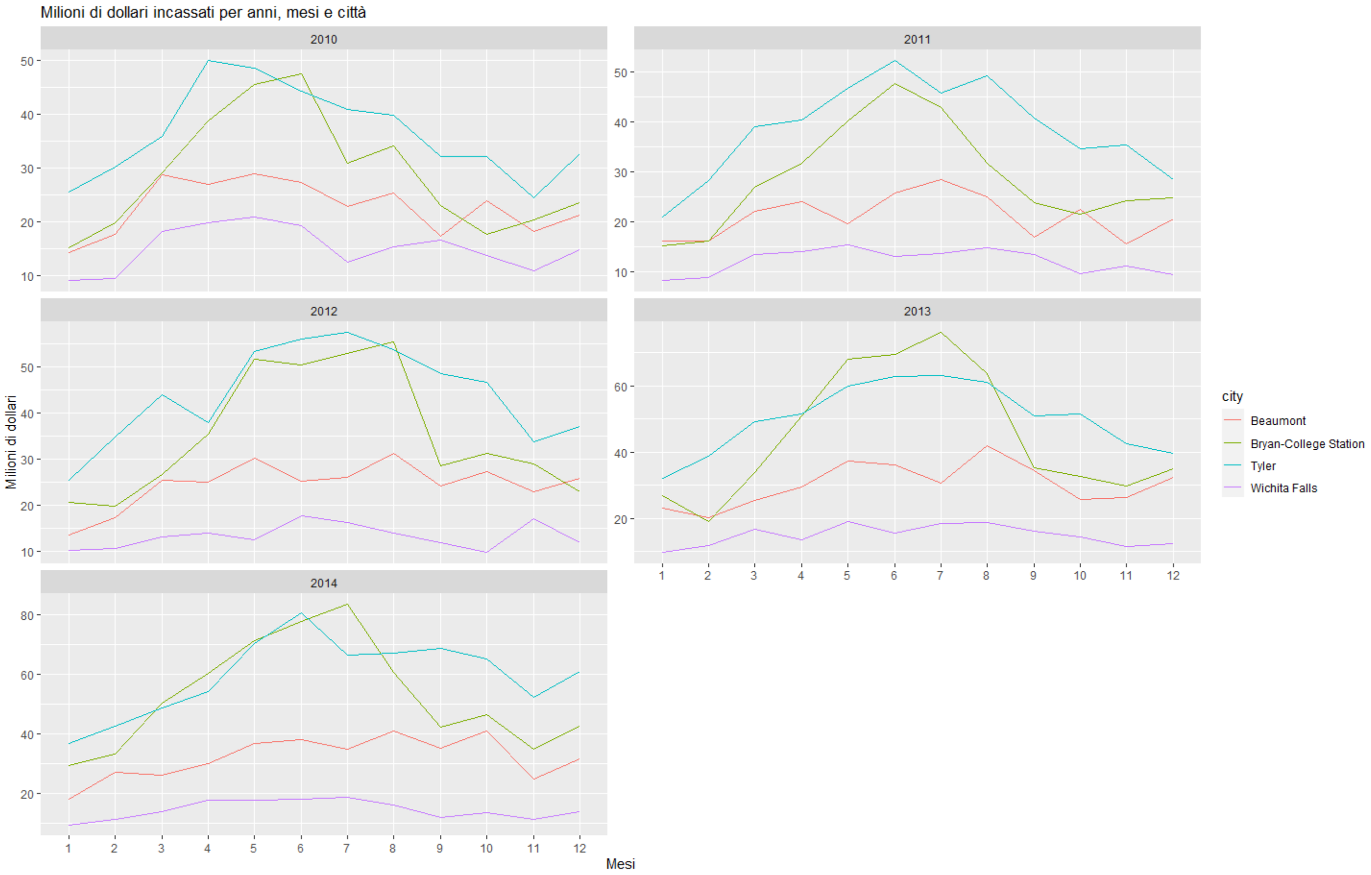
- **Beaumont**: The millions of dollars fluctuated during the years 2010 to 2014 around $20 to $40 million, peaking in 2013 with around $41 million collected from property sales in August.
- **Bryan-College Station**: The millions of dollars grossed reached 80 million in the year 2014 in the month of July. For the same month and also that of August, in the years 2012 and 2013, box office peaks were reached in the range of 50 - 65 million respectively.

- **Tyler**: The highest grossing millions of dollars were during the years 2014 and 2013, when peaks of $60 to $80 million were reached, mainly in June and July for the respective years.
- **Wichita Falls**: the millions of dollars collected were relatively low compared to all the other cities, only in 2010 were there higher takings during the months of March to June.

All in all, it can be concluded that most of the takings were made during the spring/summer periods for each city; in particular, the year 2014 was the year with the most successful takings.

This all corresponds with the analyses and conclusions drawn from the graphs showing sales for each city and historical period.

Millions of dollars collected for years, months and cities


Milioni di dollari incassati per anni, mesi e città
Milioni di dollari
Mesi
city
Beaumont
Bryan-College Station
Tyler
Wichita Falls