Mini Project 1

Hypothesis Testing:
Seasonality on Housing Price

IOD - Cohort 2023-11-14-DS-PT-AU
Huy Phan

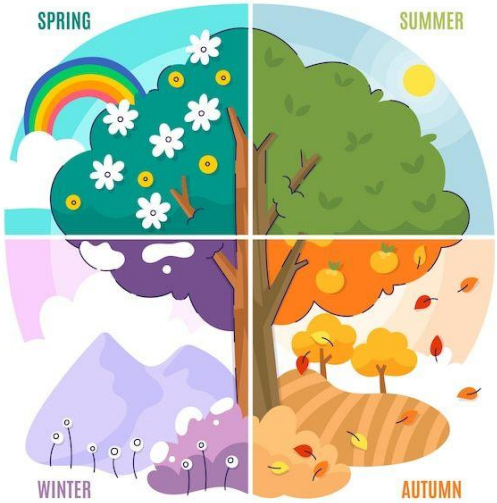
BUSINESS QUESTION

“Do seasons (the time of year) impact the average housing price?”

To answer this business question, the following hypothesis is created:

- ***Null hypothesis***: There is no significant difference in the average selling price between seasons (spring, summer, autumn, winter).
- ***Alternative hypothesis***: There is a significant difference in the average selling price between seasons

The testing technique used in this context is **one-way ANOVA test**.



DATA OVERVIEW

The data set contains housing sales data of the Sydney property market, from 13/01/2016 to 01/01/2022 (about 6 years' worth of data). (Link: https://www.kaggle.com/datasets/alexlau203/sydney-house-prices)

Important details of the data set:

- 11160 data points and 17 features
- No missing values

Key steps of initial data pre-processing:

- Data type correction
- Examination and handling of outliers
- New feature creation (adding column to present the ‘season’)
- Data shrinkage to only relevant features (‘price’ and ‘season’ for the hypothesis testing)

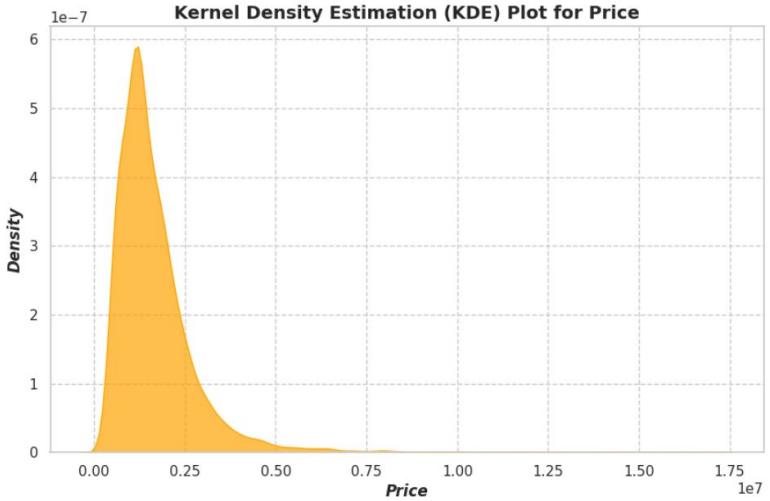
Understanding of the features

```
RangeIndex: 11160 entries, 0 to 11159
Data columns (total 17 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   price                   11160 non-null  int64
 1   date_sold               11160 non-null  object
 2   suburb                  11160 non-null  object
 3   num_bath                11160 non-null  int64
 4   num_bed                 11160 non-null  int64
 5   num_parking             11160 non-null  int64
 6   property_size           11160 non-null  int64
 7   type                    11160 non-null  object
 8   suburb_population       11160 non-null  int64
 9   suburb_median_income    11160 non-null  int64
 10  suburb_sqkm             11160 non-null  float64
 11  suburb_lat              11160 non-null  float64
 12  suburb_lng              11160 non-null  float64
 13  suburb_elevation        11160 non-null  int64
 14  cash_rate               11160 non-null  float64
 15  property_inflation_index  11160 non-null  float64
 16  km_from_cbd             11160 non-null  float64
dtypes: float64(6), int64(8), object(3)
```

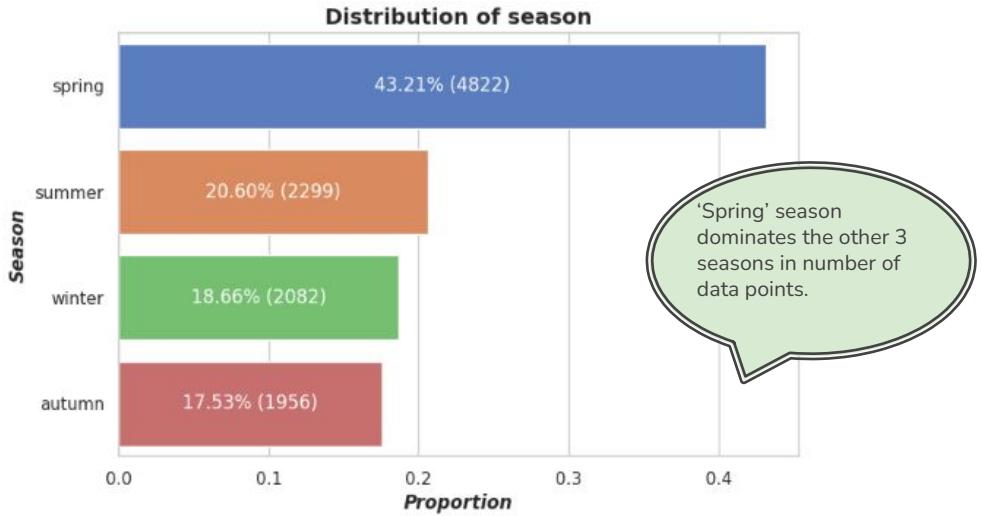
- **price**: *the price of the property in AUD*
- **date_sold**: *the date the property was sold*
- **suburb**: *the suburb the property is situated in*
- **num_bath**: *the number of bathrooms in the property*
- **num_bed**: *the number of bedrooms in the property*
- **num_parking**: *the number of parking spaces on the property*
- **property_size**: *the size of the property in square metres*
- **type**: *the type of building*
- **suburb_population**: *the population of the suburb the property is situated in*
- **suburb_median_income**: *the median income of the suburb the property is situated in*
- **suburb_lat**: *the latitude of the suburb that the property is situated in*
- **suburb_lng**: *the longitude of the suburb that the property is situated in*
- **suburb_elevation**: *the elevation of the suburb that the property is situated in.*
- **cash_rate**: *the cash rate at the time the property was sold*
- **property_inflation_index**: *the residential property price inflation index of the quarter that the property was sold in*
- **km_from_cbd**: *the distance between the property and the centre of Sydney CBD.*

Distribution of ‘price’ data

The distribution of ‘price’ is **positively skewed** and exhibits a **tendency towards normality**.



Distribution of ‘season’ data



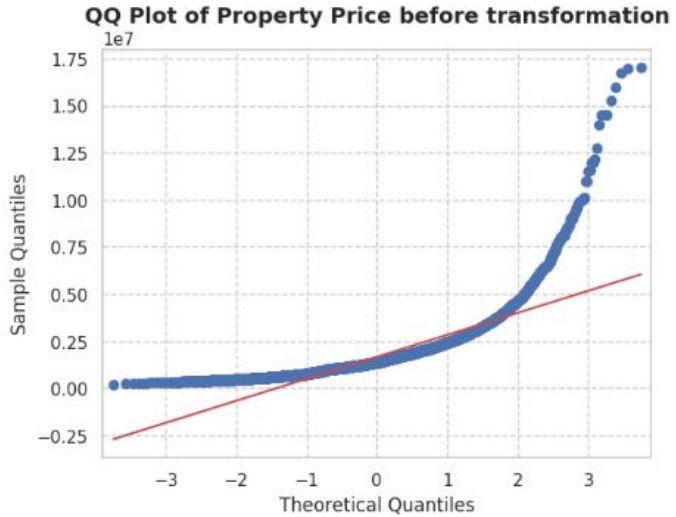
ANOVA assumption check

Validation check on the assumptions of ANOVA test:

- **Normality** – Each sample was drawn from a normally distributed population.
- **Equal Variances** – The variances of the populations that the samples come from are equal.
- **Independence** – The observations in each group are independent of each other and the observations within groups were obtained by a random sample.

Assumption 3 is true because each sale can only have 1 season value, and we assume the data was random sampled. Hence, it is worth to examine the validity of Assumption 1 & 2

Normality check (1)

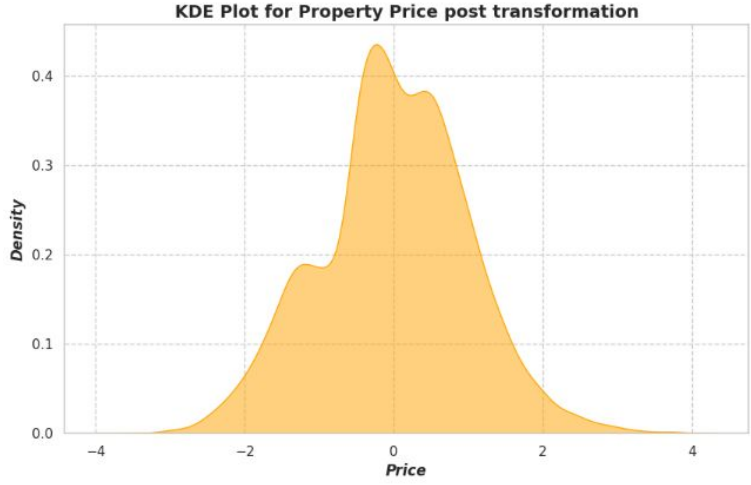


The QQ plot indicates the distribution of ‘price’ doesn’t match a normal distribution.

To make the distribution of ‘price’ more Gaussian, Yeo-Johnson transformation is performed.

Normality check (2)

After the transformation, the data is now **normally distributed** and passes the normality check.



Equal Variances check

In order to check for homogeneity of variance (all the groups have equal variance), the Levene's test of homogeneity is conducted with the following hypothesis:

- Null hypothesis: The variance is the same across the season groups.
- Alternative hypothesis: At least one season group has a different variance

The **p-value of this test = 3.20e-09**

Since the p-value < 0.05 (chosen significance level), we reject the Null Hypothesis and accept there is significant difference between the groups' variance.

Price by season distributions

Hypothesis testing outcome

Since the homogeneity of variance is violated, the use of the non-parametric Kruskal test would be more appropriate for this hypothesis testing.

p-value for Kruskal test: 5.63e-28 (which isn't much different from the p-value of one-way ANOVA test: 2.19e-28)

For this p-value, we r**eject the Null Hypothesis** (at 0.05 significance level) and accept that there is a **significant difference between the average housing price of the seasons**.

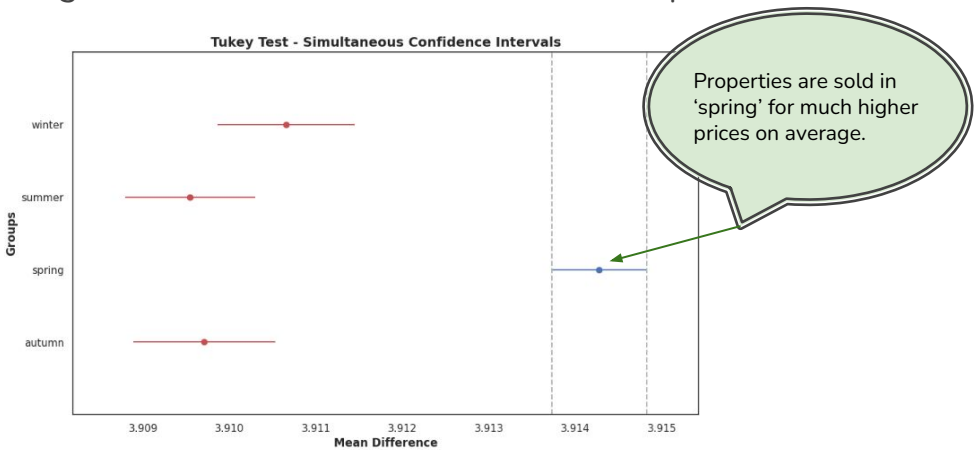
Identify the seasons that are different

The primary hypothesis testing does not give information on which particular season is different from the others in terms of average housing price.

The Tukey test is carried out to gain more insights on the difference between season pairs.

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
autumn	spring	0.0046	0.0	0.0032	0.0059	True
autumn	summer	-0.0002	0.9932	-0.0017	0.0014	False
autumn	winter	0.0009	0.4299	-0.0007	0.0026	False
spring	summer	-0.0047	0.0	-0.006	-0.0034	True
spring	winter	-0.0036	0.0	-0.005	-0.0023	True
summer	winter	0.0011	0.2513	-0.0004	0.0027	False



Bonus content

Following is the details of an extreme outlier that was removed from the data set during data preprocessing. The property was sold for $60 mil, and no other selling prices in the data set could come close to this level.

price	60000000
date_sold	3/10/21
suburb	Kurraba Point
num_bath	20
num_bed	29
num_parking	19
property_size	4240
type	House
suburb_population	1521
suburb_median_income	81744
suburb_sqkm	0.235
suburb_lat	-33.84211
suburb_lng	151.2228
suburb_elevation	26
cash_rate	0.1
property_inflation_index	197.9
km_from_cbd	2.82
date_sold_converted	2021-10-03 00:00:00

More details on the transaction if you're interested: https://www.realestate.com.au/news/waterfront-estate-on-kurraba-point-sells-to-melbourne-developer-for-60-million/