BALA301

Data-mining

SAS Software

Maxwell Wheeler

June 19th, 2023

Section A –

Question 1:

Data mining is defined as “The nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data stored in structured databases” Fayyad et al. (1996) Data mining is the process of discovering knowledge from large existing amounts of data. There are many different types of data mining such as Pattern mining, knowledge mining, and insight mining, decision trees, artificial neural networks, and SVMs. These processes involve various statistical, mathematical, and artificial intelligence techniques and algorithms to extract patterns in data. Such patterns could be affiliations, relations, groupings, trends, and predictions. The Data mining processes generally follow these 6 steps:

Step 1: Business Understanding

During the business understanding step a data mining project beings by understanding the business problem or objective. This is where goals, requirements, and constraints of the project are established. An example of this would be establishment of the goal predicting box-office receipts (i.e., financial success) of a movie before its release. Some constraints would be to mitigate the risk associated with the movie production budget and different decision making in the entertainment industry.

Step 2: Data Understanding

Data understanding involves acquiring and understanding the data thoroughly. An example in context would be obtaining data related to previous movies, box-office revenues, production budgets, release dates, marketing strategies, and audience demographics, and researching these various topics to gain a strong understanding of them.

Step 3: Data Preparation

Data preparation involves gathering all data needed and preparing it for analysis. It also involves cleaning data for missing values and transforming different variables if needed to.

An example of data preparation in context would be filtering data of previous movies into specific genres or time periods. And filtering data to gain better insight into the production costs of films and the revenue produced within the first 5 years of release.

Step 4: Model Building

Model building consists of building predictive models based of prepared data. Different analysis techniques such as regression analysis, decision trees, and AI algorithms can be used to predict box-office statistics prior to release based on various factors. These models could contain data regarding production budgets, marketing budgets, market competition, movie genres, and other relevant variables to forecast predictions of sales and success.

Step 5: Testing and Evaluation

Testing and evaluation involves taking models built in step 4 and are evaluated carefully to find out their predictive performance. The testing and evaluation process is done by evaluating metrics such as R^2, mean squared error and using them to measure predictive performance. An example of this would be testing and evaluating models of previous box-office revenues and forecast accurately what future revenues could potentially be.

Step 6: Deployment

The last step in the data mining process is the deployment of such models formulated. Once the models are at a satisfactory confidence in accuracy they can be deployed for results. The deployment of the models can be used by investors, distributors, production studios, and many others to make informed decisions about movie release dates, marketing strategies, and revenue predictions. Note that Steps 1-3 account for 85% of total project time that the above 6 steps are highly repetitive and experimental.

Question 2:

In Table I the results display a multiple regression analysis, and it is evident that the regression model is significant. We know this due to the (F statistic = 31.89, p value < 0.05), indicating that the independent variables (belief in fate and belief in fortune-tellers) have a significant impact on the brand logo sensitivity, supporting our alternate hypothesis H_1.

H_0: The belief in fate and belief in fortune-tellers will not have a significant influence on the brand logo sensitivity.

H_1: The belief in fate and belief in fortune-tellers will have a significant influence on the brand logo sensitivity.

Table II shows the results of a collinearity statistics test, and display that belief in fate, belief in fortune tellers, belief in magic and fictional figures, belief in lucky charms, and belief in superstitious rituals all hold significant beta coefficients (p value < 0.05). This suggests that each individual variable contributes to the prediction of brand logo sensitivity, supporting our alternate hypothesis H_2.

H_0: The belief in magic and fictional figures, belief in urban legends, belief in lucky charms, belief in superstitious rituals will not have a significant influence on brand logo sensitivity.

H_2: The belief in magic and fictional figures, belief in urban legends, belief in lucky charms, belief in superstitious rituals will have a significant influence on brand logo sensitivity.

In Table III we are presented with the coefficient of determination (R^2) which will tell us how much of the independent variables explain the dependent variable. It was found that all independent variables explain 40% of the variance in brand logo sensitivity which suggests a

moderate level of predicting power of the brand logo sensitivity. Thus, we accept our alternate hypothesis H_3.

H_0: The combined influence the independent variable superstitious beliefs will significantly explain the variation found in brand logo sensitivity.

H_3: The combined influence the independent variable superstitious beliefs will significantly explain the variation found in brand logo sensitivity.

All in all, the results tell us that superstitious beliefs have significant influence on consumers' information processing when evaluating different brand logos. Specifically, belief in fate, belief in fortune tellers, belief in magic and fictional figures, belief in lucky charms, and belief in superstitious rituals all play a role in shaping the consumers' sensitivity to brand logos. It can be confidently said that the combines effects of these beliefs can explain a significant portion of the variation in brand logo sensitivity.

Section B –

Question 1:

Descriptive Statistics of 300 houses (Mean, Variance and Range (min - max), and Shape (skewness and kurtosis))

Cluster #1 SIZE - *Appendix 1*

Living Area SqFt: (1,130.74, 54,125.74 and 334 – 1,500, -0.3905 and -0.3328)

Basement SqFt: (882.31, 129,444.50 and 0 – 1645, -0.5477 and 0.1374)

Garage SqFt: (369.45, 31,065.15 and 0 – 902, -0.3738 and 0.0392)

Decks and Porches SqFt: (118.26, 17,585.86 and 0 – 897, 1.4534 and 3.6796)

Lot Size SqFt: (8,294.14, 11,047,565.79 and 1,495 – 26,142, 1.0093 and 4.5758)

Cluster #2 DATE - *Appendix 2*

Year Sold: (2008, 1.72 and 2006 - 2010, 0.0521 and -1.11656)

Month Sold: (5.91, 6.35 and 1-12, 0.2196 and -0.1661)

Age of House When Sold (years): (45.89, 745.98 and 1-135 years old, 0.1953 and -0.5449)

Cluster #3 BED BATH - *Appendix 3*

Bedrooms: (2.51, 0.48 and 0 – 4, -0.7203 and 0.5582)

Fireplaces: (0.39, 0.32 and 0 – 2, 1.1144 and 0.2571)

Full Bath: (1.68, 0.44 and 1 – 4, 0.5397 and -0.3946)

Half Bath: (0.25, 0.20 and 0 – 2, 1.3819 and 0.5084)

Cluster #4 OVERALL - *Appendix 4*

Overall Condition of The House: (5.8, 1.42 and 3 – 9, 0.4044 and 0.0716)

Overall Material and Finish of the house: (5.45, 1.44 and 1 – 9, -0.3144 and 1.0929)

The relationship between Clusters and “Sale price in dollars.”

Cluster #1 SIZE / 3 Cluster IDs

Cluster ID 3 holds the highest sales price in dollars (170,000+) due to majority of houses having the highest area of decks (50 – 525sqft), basement (550 – 1,750sqft), lot size (1,700 – 17,750sqft), and garage (220 – 735sqft).

Cluster ID 1 has the lowest sales price in dollars ($103,800) due to majority of houses having the lowest area of decks (0 – 225sqft), basement (0 – 1275sqft), lot size (1,500 – 12,000sqft), and garage (0 – 530sqft).

Cluster #2 DATE / 3 Cluster IDs

Cluster ID 2 holds the highest sales price in dollars ($149500) due to majority of houses having the highest year sold 2008 - 2010, majority of houses having low ages of 0 – 75 years old, and majority of houses sold in the months Jan - July.

Cluster ID 1 holds the lowest sales price in dollars ($122,500) due to majority of houses being sold in years 2006 - 2008, being the eldest at 25 – 135 years old, and majority of houses sold in the months Feb – Aug.

Cluster #3 BED BATH / 2 Cluster IDs

Cluster ID 2 holds the highest sales price in dollars ($157,000) due to the houses having a majority of 2-4 bedrooms, 0-1 half bathrooms, 1-4 full bathrooms, and 0-2 fireplaces.

Cluster ID 1 holds the lowest in sales price in dollars ($117,500) due to having the lowest majority of 1-3 bedrooms, 0-1 half bathrooms, 1-2 full bathrooms, and 0-1 fireplaces.

Cluster #4 OVERALL / 2 Cluster IDs

Cluster ID 1 holds the highest sales price in dollars ($144,500) due to the houses in this cluster having a majority overall condition score of 6-9, and an overall material and finish score of 5.5 – 9.5.

Cluster ID 2 holds the lowest in sales price in dollars ($130,500) due to majority of houses having an overall condition score of 3-6, and an overall material and finish score of 1-8.

Question 2: " { } = code "

The code { AVG_LA_TEMP.head(5)["AVG_TEMP"].mean } was run to obtain an average mean of LA's temperature for the first 5 years of the dataset. Years (1849-1853). The mean is 15.63.

The code { AVG_LA_TEMP.tail(5)["AVG_TEMP"].mean } was run to get an average mean of the temperature for the last 5 years of the dataset. Years (2009-2013). The mean is 16.73.

The code { AVG_LA_TEMP.tail(60).describe() } was run to determine if there are any significant differences in the average temperature over 60 years. By running such code, we are presented with minimum and maximum average temperature of the last 60 years (15.12 and 18.12). With a 3-degree difference we can say that this is a significant average temperature change that has occurred over the last 60 years.

To calculate the coefficient of variation for the entire dataset the code {coefficient_of_variation = AVG_LA_TEMP["AVG_TEMP"].std() / AVG_LA_TEMP["AVG_TEMP"].mean } was run to set the calculation needed to determine the coefficient of variation (Standard Deviation / Mean). { print (coefficient_of_variation) } was run to give us the coefficient of variation which is 0.0358.

To forecast the average temperature for the year 2024 code { linear_regression.slope * 2024 + linear_regression.intercept } was run to forecast the average temperature for the year 2024 which was found to be 16.36 degrees celsius. To find the slope and intercept the following code was run prior to the above code. { linear_regression = stats.linregress(x=AVG_LA_TEMP.YEAR, y=AVG_LA_TEMP.AVG_TEMP) } this will set our equation for the linear_regression. { linear_regression.slope } and { linear_regression.intercept } gives us both the slope and intercept for the dataset (0.01 and 6.01).

Analysis of LA's historical temperature-centric data could lead to the discovery of different patterns and trends which could in hand open opportunity for potential business opportunities. Such potential business opportunities could be increased efficiency of the agriculture industry, capitalization of the tourism industry, and new renewable energy sources. As temperature holds a significant role in crop growth and annual yield, understanding temperature data can provide insights on what crops will grow the fastest and biggest during different times of the year. Regarding tourism, temperature also plays a crucial role in when such tourist activities are undertaken and engaged in. Organizing events, promoting different seasonal activities, and planning when different businesses will be popping off can all be taken advantage of as temperature centric data can reveal when the best time to do so is. Not only can the data give insight into when to plan events in LA, but it can also open opportunity for new energy efficient solutions. This can be done by identifying period of high energy consumption, peak usage, and low usage, and leverage these statistics to their advantage by implementing smarter temperature management, smarter thermostats, and offering energy conserving solutions to various businesses and industries in LA.

```
In [1]: import pandas as pd
In [2]: AVG_LA_TEMP = pd.read_csv('https://raw.githubusercontent.com/jpadillo/Analyzing-Weather-Dataset/master/LA.csv')
In [3]: AVG_LA_TEMP.columns = ['YEAR', 'CITY', 'COUNTRY', 'AVG_TEMP']
In [4]: AVG_LA_TEMP.head(5)["AVG_TEMP"].mean()
Out[4]: 15.680000000000001
In [5]: AVG_LA_TEMP.tail(5)["AVG_TEMP"].mean()
Out[5]: 16.73
In [6]: AVG_LA_TEMP.tail(60).describe()
```

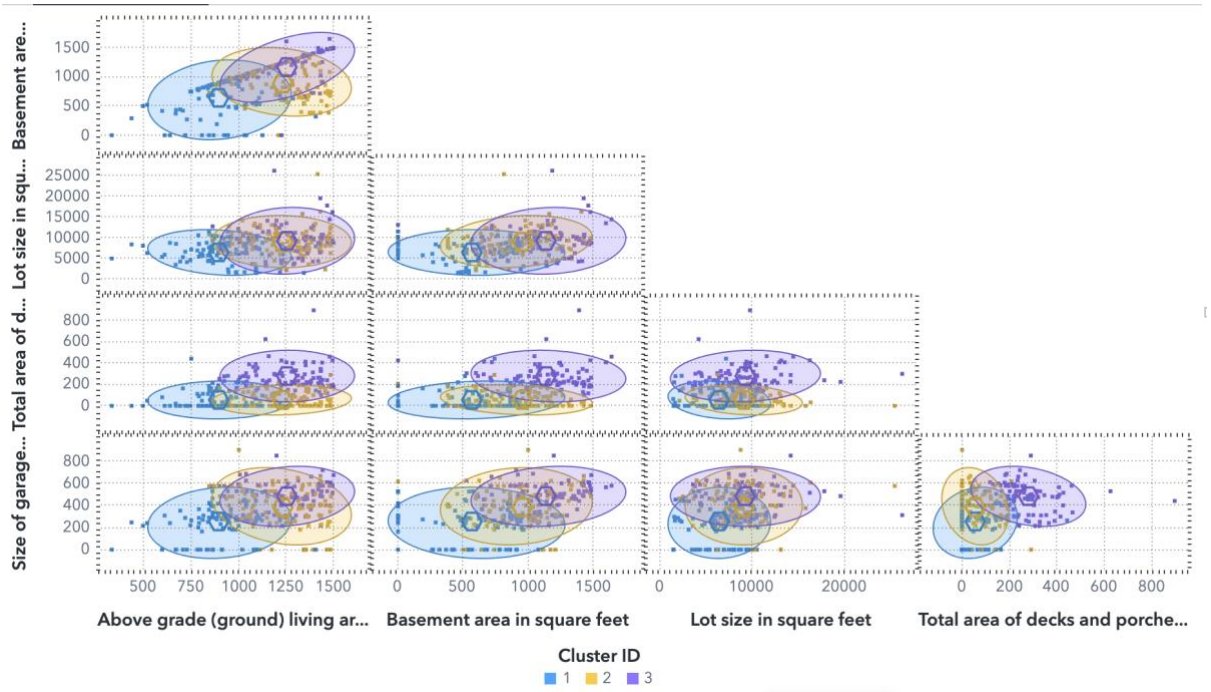
Out[6]:

	YEAR	AVG_TEMP
count	60.000000	60.000000
mean	1983.500000	16.216833
std	17.464249	0.565315
min	1954.000000	15.120000
25%	1968.750000	15.877500
50%	1983.500000	16.205000
75%	1998.250000	16.625000
max	2013.000000	18.120000

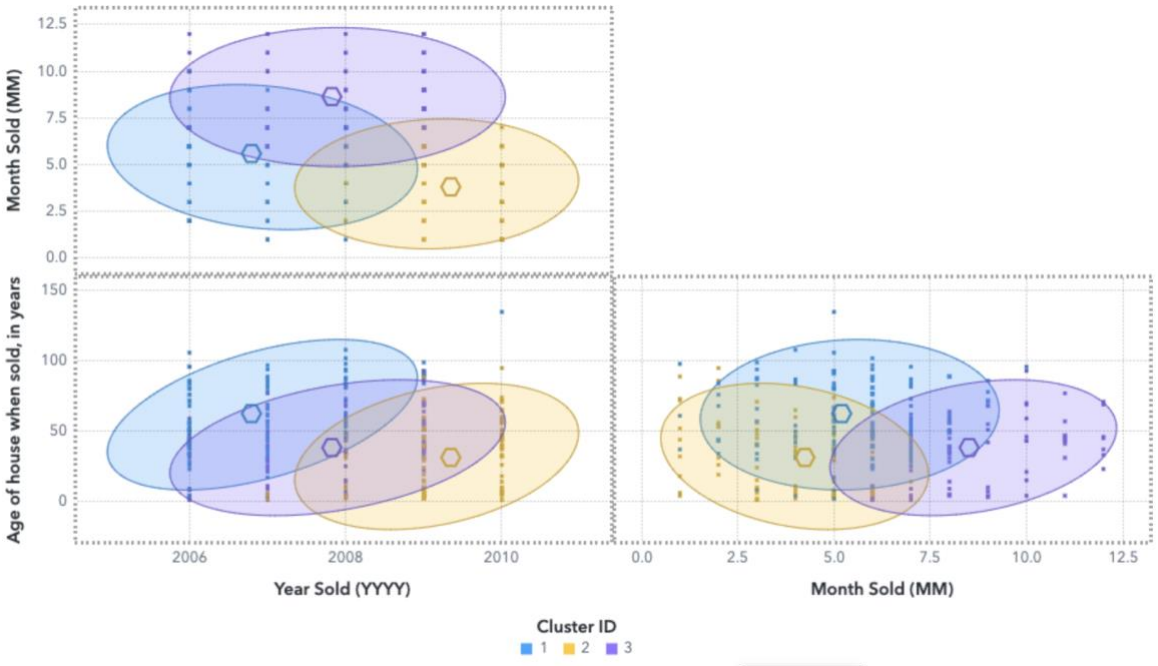
```
In [7]: from scipy import stats
In [8]: coefficient_of_variation = AVG_LA_TEMP["AVG_TEMP"].std() / AVG_LA_TEMP["AVG_TEMP"].mean()
In [9]: print (coefficient_of_variation)
0.03581313896530716
In [10]: linear_regression = stats.linregress(x=AVG_LA_TEMP.YEAR, y=AVG_LA_TEMP.AVG_TEMP)
In [11]: linear_regression.slope
Out[11]: 0.005113830043010179
In [12]: linear_regression.intercept
Out[12]: 6.006345702098859
In [13]: linear_regression.slope * 2024 + linear_regression.intercept
Out[13]: 16.35673770915146
```

Appendix:

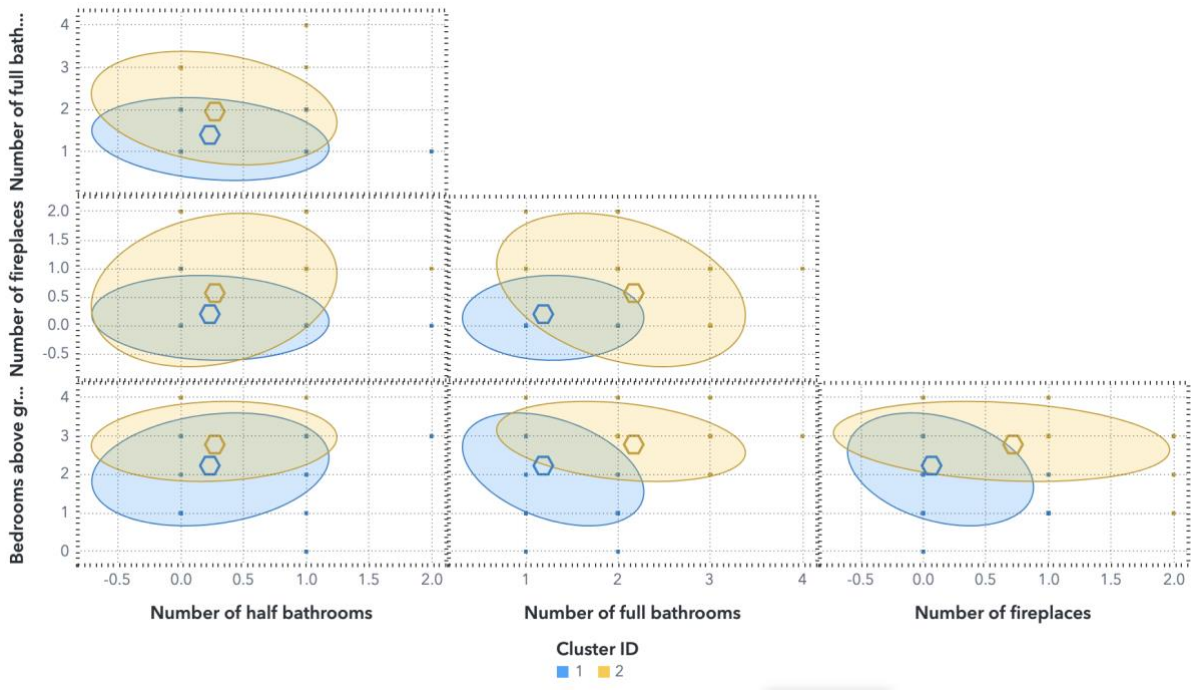
1.



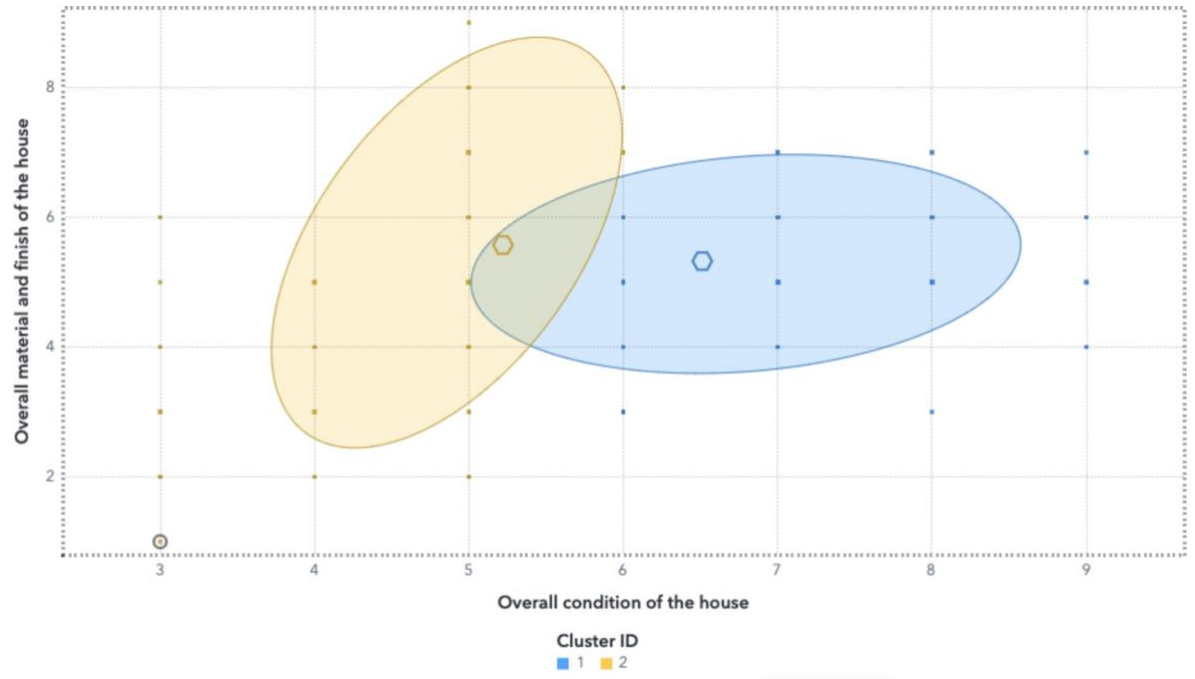
2.



3.


Number of full bath...
Number of fireplaces
Bedrooms above gr...
Number of half bathrooms
Number of full bathrooms
Number of fireplaces
Cluster ID
1
2

4.


Overall material and finish of the house
Overall condition of the house
Cluster ID
1
2