Jordan Black

CGS2020 202501

Professor Lakshmi Prayaga

3/31/2025

**Assignment 6 Regression and K-means Report Summary**

**Complete the following**

**Part 1**

**1. Regression assignment. The goal of this assignment is to look at the  money spent on marketing via facebook, youtube and newspaper and its effect on sales.**

**a. Open the RegressionAssignment2.r file. Check the libraries in the file and make sure to install any packages you may need.**

**b. The data file you will be using is "marketing". It is included in the package called datarium. So install that package as well**

**c. The rest of the file is similar to what we did in the regression module, running the r program and looking at the relationship between facebook marketing $ and impact on sales.**

**d. Run the example file and answer the following:**

**e. Given marketing through facebook and sales, which is the dependent variable, which is the independent variable?**

**Independent Variable: Facebook (Marketing budget spent on Facebook)**

**Dependent variable: Sales (resulting sales values)**

**f. State the null hypothesis and alternate hypothesis as we did in the income vs. happiness example. Remember null hypothesis is always that there will be no effect of one variable on the other, and the alternate hypothesis is that there will be an influence of one variable on another.**

**Null Hypothesis: (H0) Facebook marketing has no significant effect on sales.**

**Alternative Hypothesis: (H1) Facebook marketing Does have a significant effect on sales.**

**g. Run the example and interpret the results referencing the r squared value and p value. What do they tell you about the accuracy of your regression model?**

**A computer screen shot of a error

AI-generated content may be incorrect.**

**Slope: = 0.2025: For every $1 increase in Facebook marketing, sales increase by about $0.20**

**R^2 = 0.32: Approximately 33.2% of the variability in sales can be attributed to Facebook spending. This is a moderate correlation.**

**P-value <0.05 (very small): The relationship is statistically significant – strong evidence that Facebook spending impacts sales.**

**Repeat steps a to g by changing Facebook to YouTube**

**YouTube Results:**

**Independent Variable: YouTube**

**Dependent Variable: Sales**

**Null Hypothesis (H0): YouTube marketing does not affect sales.**

**Alternate Hypothesis (H1): YouTube marketing has a significant effect on sales.**

**Regression Output:**

**Slope: 0.0475**

**R^2: 0.6119**

**p-value: <2.2e-16**

**YouTube explains 61% of the variation in sales, making it the most accurate model. However, the effect per dollar spent is less significant than that of Facebook.**

**A screenshot of a computer

AI-generated content may be incorrect.**

**Repeat steps a to g by changing YouTube to newspaper**

**Newspaper Results:**

**Independent Variable: Newspaper**

**Dependent Variable: Sales**

**Null Hypothesis: Newspaper marketing has no significant effect on sales.**

**Alternative Hypothesis: Newspaper marketing has a significant effect on sales.**

**Regression output:**

**Slope: 0.0547**

**R^2: 0.052**

**p-value:0.00115**

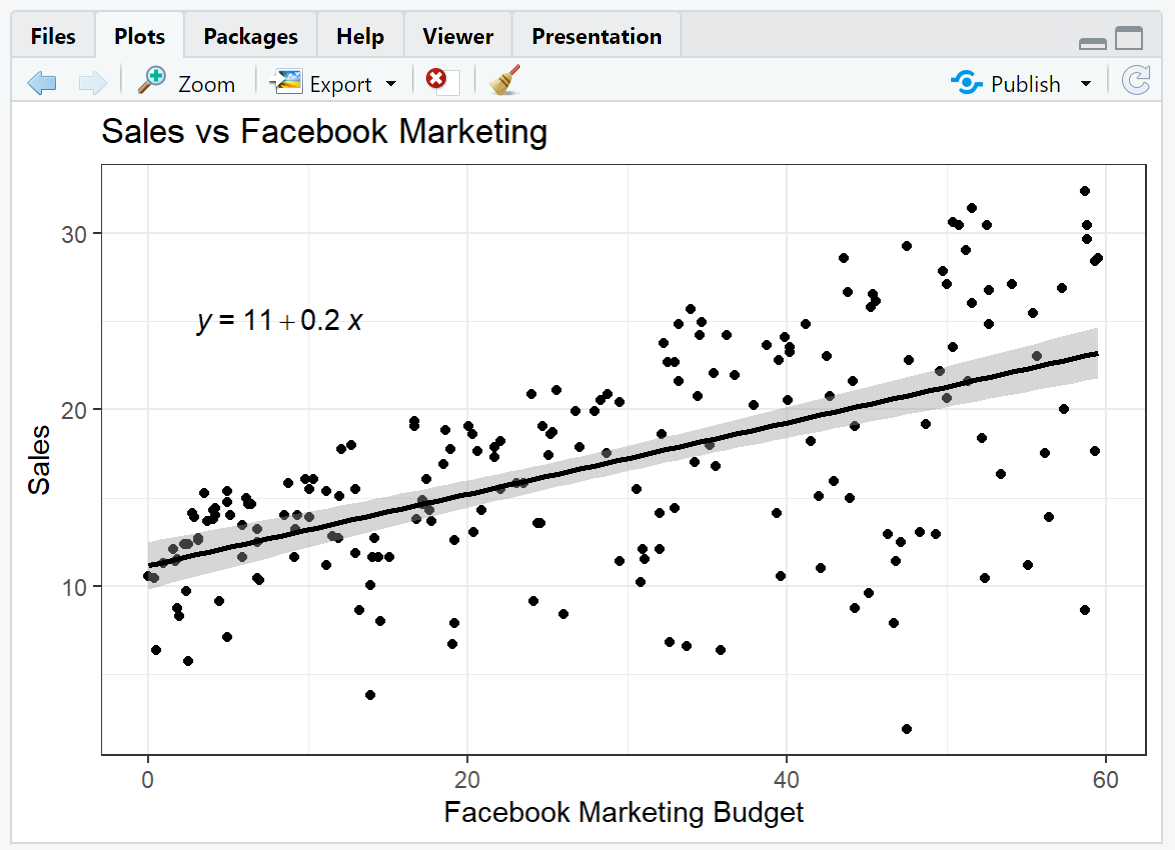
**Interpretation: The Relationship is statistically significant, but the model explains only 5.2% of the variation in sales. It is the weakest predictor overall.**

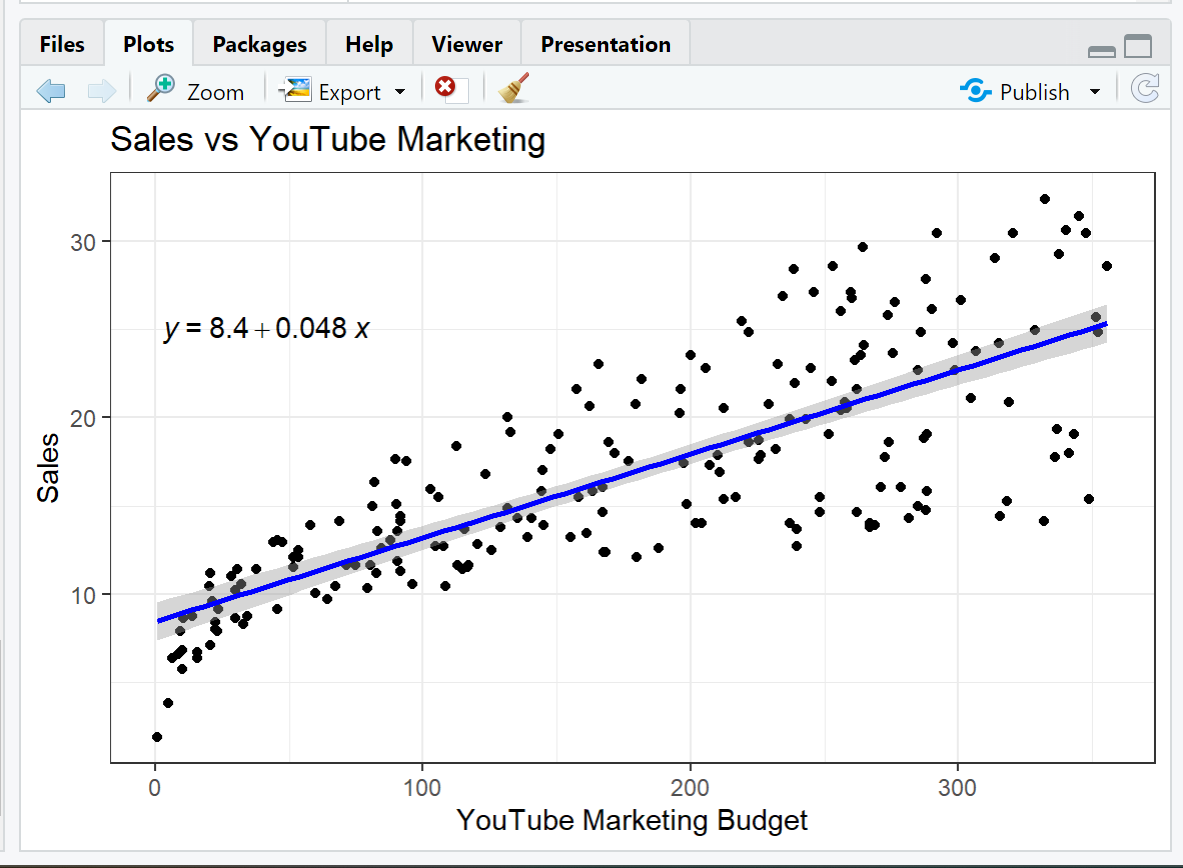
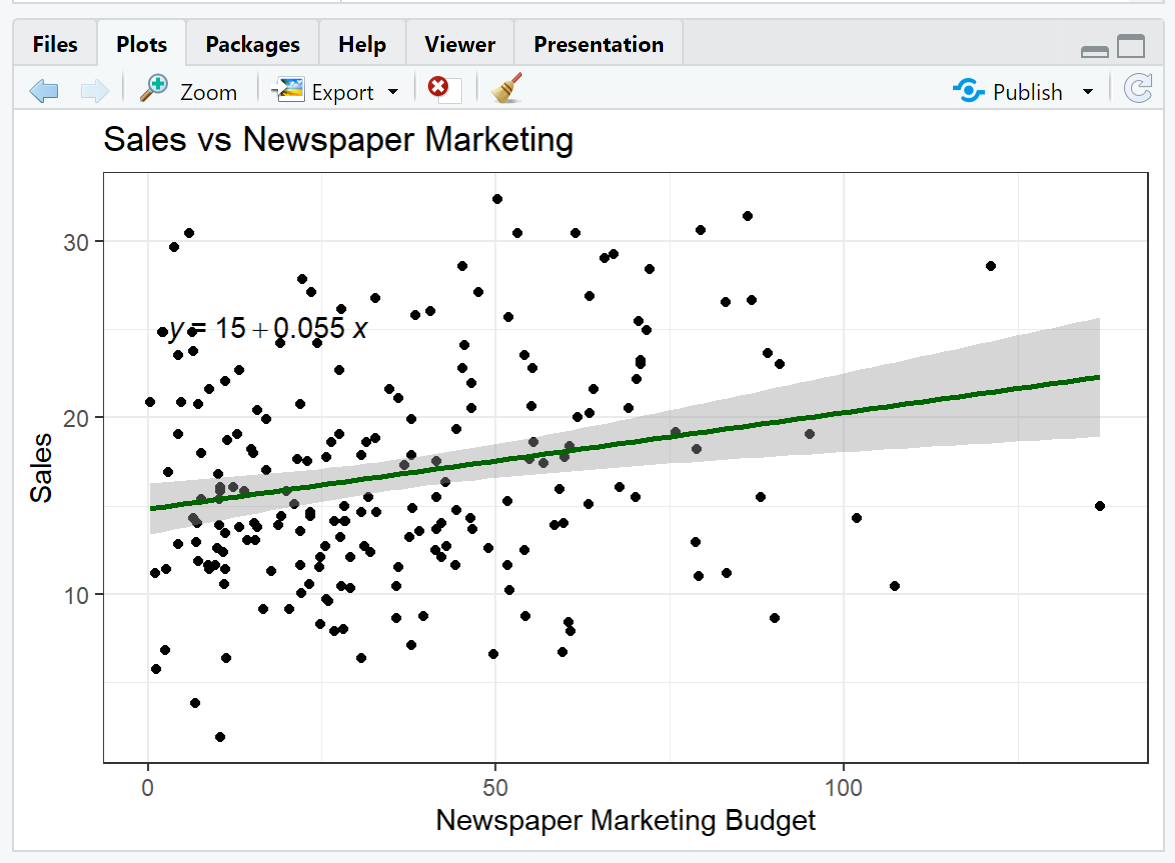
**A screenshot of a computer error

AI-generated content may be incorrect.**

**Summarize your findings and state which of the three media of marketing results in the best sales? Look at the slope of the graphs and decide which of the three venues gave the best sales per dollar.**

**Screenshot of each:**



**Medium | Slope | R^2 | p-value | Summary**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Facebook** | **0.2025** | **0.332** | **< 2.2e-16** | **The strong effect per dollar spent; model explains about 33% of the variation in sales.** |
| **YouTube** | **0.0475** | **0.6119** | **<2.2e-16** | **Best fit overall: it explains over 61% of sales variation, indicating extreme predictive power.** |
| **Newspaper** | **0.0547** | **0.052** | **<0.00115** | **Weak model: only 5.2% of sales variation is explained, and it is not a reliable predictor despite being statistically significant.** |

**Best Sales per Dollar:** The marketing return rate on Facebook shows a 0.2025 slope, indicating that each $1 investment generates $0.20 more in sales. Facebook delivers the best sales performance for each dollar spent when compared to the other two media types. Facebook provides the highest return on investment per dollar spent even though its sales variance explanation is less than YouTube's (represented by a lower R²).  
  
**Most Accurate Model:** YouTube features the largest R² value of 0.6119, which means the regression model explains 61% of the variation in sales based on YouTube marketing investments. The model demonstrates strong predictive capabilities with an excellent fit despite having a smaller slope compared to Facebook. This model demonstrates superior statistical strength for comprehensive predictive accuracy.  
  
**Least Effective:** The R² value of 0.052 indicates that newspaper advertising accounts for only 5.2% of the variations in sales despite its statistically significant p-value and slope of 0.0547. The low R² value demonstrates that newspaper marketing is an unreliable predictor of sales, despite a positive slope, and should not be relied upon.

**Part 2 Of Assignment:**

**For this assignment you will use the files - KmeansEuroJobVAssignment. and Eurojob.csv**

**Then open the r file and import the Eurojob.csv.**

**1. Run each line of the code and observe the output.**

**In this example I have used Man (manufacturing)  and SPS (Services) as the two variables**

**Print the 2 plots from this file. Notice that the second plot, automatically labels x and y axes as Dim1 and Dim2. This is a default output given by Fviz. It basically takes all the variables in the data set and chooses the best as the two new variables. You will learn more about this technique called principal component analysis when you do more advanced courses. For now this is sufficient to know.**

**2. Next play with a combination of any other two variables from the Euro jobs and see what kind of clusters you get. For example with MAN and SPS, you'll notice there is more data for one cluster vs. the other. See if this changes and document it.**

**Step 1: Clustering with Manufacturing (Man) And Services (SPS)**

**After loading the Eurojob.CSV file and scaling the numeric variables, I arna k-means clustering using two variables: (Man) and (SPS). As you did for your analysis, you then set the number of clusters to 2.**

**A screenshot of a computer screen

AI-generated content may be incorrect.**

**Observations:**

**The cluster plot demonstrated a distinct division between two distinct economic profiles.**

**The first cluster comprised nations that exhibited elevated service sector employment characteristics typical of post-industrial or service-oriented economies. Countries in Cluster 2 exhibited higher levels of employment in the manufacturing industry. The majority of European countries fell within Cluster 1, indicating a shift towards service-based employment across the continent.**

**PCA Cluster Plot Using All Variables  
The fviz\_cluster() function helped me create a cluster plot based on PCA analysis. This function compresses high-dimensional employment data into two dimensions: The dimensions Dim1 and Dim2 demonstrate the natural grouping of countries when considering data from all sectors.** **A screen shot of a graph

AI-generated content may be incorrect.  
  
 PCA Cluster Plot (All Variables):  
A majority of countries form a tight group in Cluster 1, indicating their employment patterns share common traits.  
  
Turkey and Yugoslavia represent a separate group in Cluster 2 which indicates they follow different economic systems.  
  
The dimensionality reduction technique clearly demonstrates the existing group structures.**

**A screen shot of a graph

AI-generated content may be incorrect.**

**Step 2: Clustering with Finance (Fin) and Transport (TC)**

**I executed the same k-means process using Finance (Fin) and Transport (TC) data.**

**Clusters using Finance and Transport:**

A screenshot of a computer screen

AI-generated content may be incorrect.

**The clusters are less clearly separated.**

**Employment levels in finance and transport industries are comparable across most countries.**

**The presence of a small number of outliers reduces the effectiveness of this pair of variables for clustering tasks.**

**The PCA Plot demonstrates Finance and Transport as variables within the broader analysis.**

**Final Summary:**

A screenshot of a computer

AI-generated content may be incorrect.

**Man + SPS demonstrated clusters which displayed greater meaning and visibility by showing actual economic variations.  
  
The PCA visualization verified the cluster structure by analyzing all variables.  
  
The combination of Fin and TC variables demonstrated restricted ability to create clusters which confirms the need to choose impactful variables for unsupervised analysis.**