#### **Part One—Identify a Focal Point and a Dependent Variable**

# Target of Analysis:

The Higher Management is extremely critical related to Cost that is spent in the Business Intelligence program what I am working right now. So, the priority is to control and have a clear forecast on the amount of budget we are going to spend related to this program until the end. So as per that the variable of interest for me in this project is “**Cost Per Month related to Cloud Infrastructure**” and predicting it for the duration of time which will last for another 3 years. There are multiple independent variables that are related to / has impact on variable of interest i.e. “**Cost Per Month related to Cloud Infrastructure**”.

1. Identify the dependent variable that will guide your prediction or decision.

Dependent variable: **Cost Per Month related to Cloud Infrastructure**

1. Identify at least three independent variables that you believe have association with the dependent variable. For each independent variable, identify it as quantitative or categorical and discuss its expected impact on the dependent variable.

|  |  |  |
| --- | --- | --- |
| **Independent Variable** | | |
| **Summary of independent variable** | **Categorical or quantitative?** | **Argument for / description of the associates with the dependent variable** |
| Total Number of Tables to be Implemented | Quantitative | In this BI related Development Program we have multiple product development planned every month.  As part of this development, Number of Tables implemented in the Databases will increase every month.  This will have an impact on Sizing and Cost |
| Total Amount of Data Loaded into Tables | Quantitative | Every month Tables count will increase as part of new Development.  This will result in increase of data that is loaded into Cloud.  This will have an impact on Sizing and Cost |
| Size of Cloud Services leveraged | Quantitative | The above factors will impact the size of Cloud Services we leverage so that the performance is optimal.  As the number of tables and data increase We have to increase the size of Cloud Service so that it performs to the expected limits. |

**Part Two—Map Decisions to Outcomes**

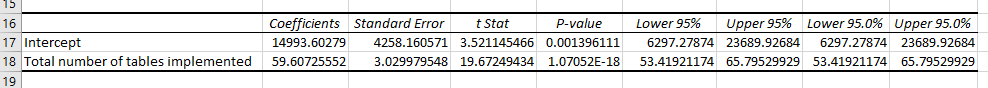
1. **For each** of the independent variables listed in part 1:
   1. Paste a screenshot of a scatterplot here (include a best fit line for all quantitative variables). Make sure the independent variable is on the horizontal axis and the dependent variable is on the vertical axis.
   2. Write the regression equation of the best fit line in the table below.

**Candidate Independent variables**

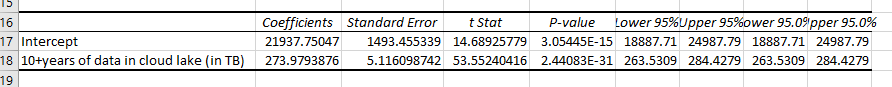
|  |  |  |
| --- | --- | --- |
| **Independent variable** | **Regression equation** | **Screenshot of scatterplot** |
| Total Number of Tables to be Implemented | Y=59.607X+1499  R-Square = 0.9280 |  |
| Total Amount of Data Loaded into Tables | Y=273.98X+21938  R-Square = 0.9896 |  |
| Size of Cloud Services leveraged | y = 24.743x – 10710  R-Square = 0.9952 |  |

**For each** of the independent variables listed in part 1, use plain language to explain why the relationship shown in the scatterplot and regression equation make sense in the context of the situation you’re exploring. However, if either the slope or intercept of the equation seems counterintuitive given your intuition, also make a note of that in your explanation. *(50-100 words per variable)*

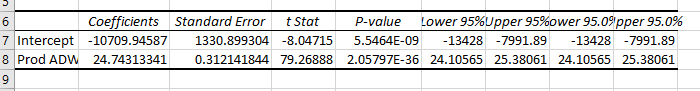
**Table vs Cost :** Independent Variable Total number of Tables to be Implemented vs Total Cost has a strong R-Square value of 0.9280.The P-Value was also very low as seen in the screen shot. The above 2 points proves that Independent variable Number of Tables has a strong relationship with dependent variable “Cost”



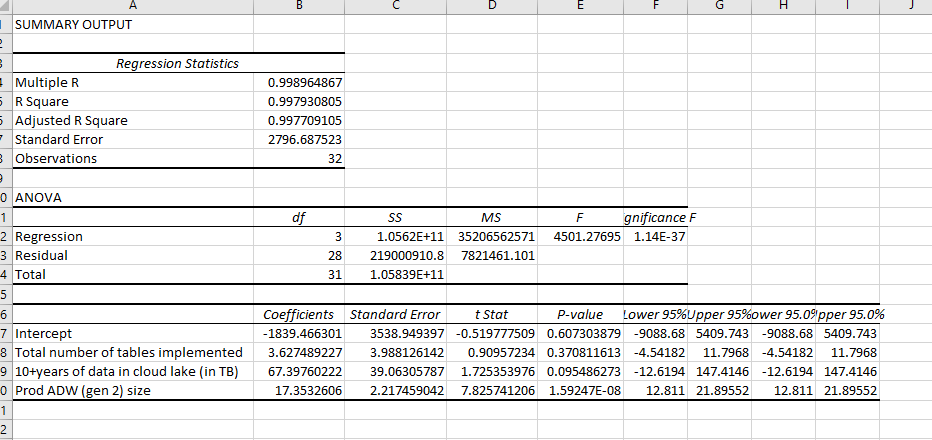
**Data vs Cost:** Independent Variable Total Amount of Data Loaded into Tables vs Total Cost has a Strong R-Square value of 0.9896. The P-Value was also very low as seen in the screen shot. The above 2 points proves that Independent variable Total Amount of Data Loaded has a strong relationship with dependent variable “Cost”.



Size vs Cost: Independent Variable “Size of Cloud Service Leveraged” vs Total Cost has a strong R-Square value of 0.9952. The P-Value was also very low as seen in the screen shot. The above 2 points proves that Independent variable Size of Cloud Service has a strong relationship with dependent variable “Cost”.



**Run a multiple regression for your data. Paste a screenshot of the results below.**



**Part ThreeGenerate a Revised Regression Equation**

**For each** of the independent variables listed in part 1, paste a screenshot of its residual plot against the dependent variable Y here in your project document

Table vs Cost:

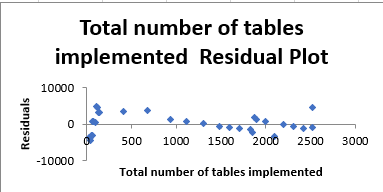
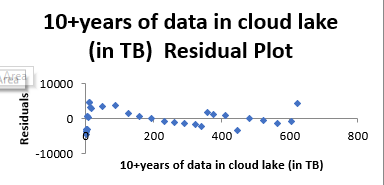


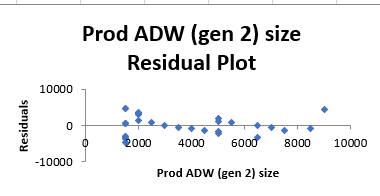
Table vs Cost shows a random pattern, indicating a good fit for a linear model.

Data vs Cost:



Data vs Cost shows a random pattern, indicating a good fit for a linear model.

Size vs Cost:



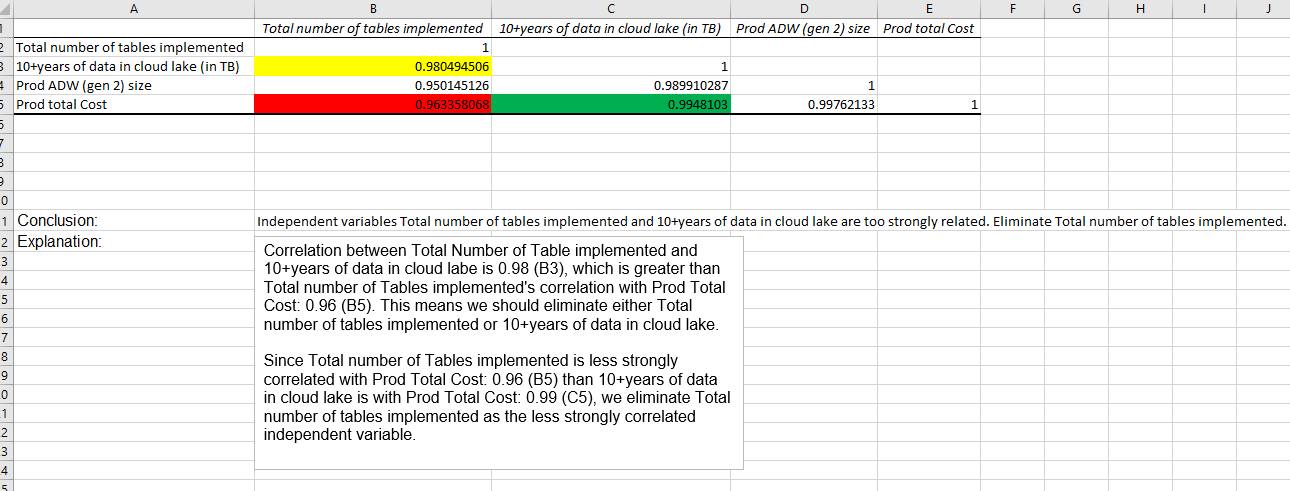
Size vs Cost shows a random pattern, indicating a good fit for a linear model.

**Addressing Multicollinearity**

1. Create a correlation table for your independent variables. Paste a screenshot of your correlation table here.

Correlation between **Total number of tables implemented** and **10+years of data in cloud lake** is 0.98(B3), which is greater than **Total number of Tables implemented’s** correlation with **Prod Total Cost** 0.96(B5). This means we should eliminate either **Total number of tables implemented** or **10+years of data in cloud lake**.

Since, **total number of tables implemented** is less strongly correlated with **Prod Total Cost** 0.96(B5) than **10+ years of data in cloud lake** is with **Prod Total Cost** 0.99(C5), we eliminate **Total number of tables implemented** as the less strongly correlated independent variable.



**Part Four—Validate Your Model**

In this project we want to build a model which will help in forecasting the Cost of Cloud Infrastructure every month with minimal to no uncertainty. So, the variable of interest is “Total Prod Cost”.

The Regression model what is build after eliminating multicollinearity has a very strong R-Squared value of 0.9978 which explain that our model has very minimal uncertainty.

Also, the P-Value is less than 5% informing that the independent variables what we have picked for forecasting has strong relationship with independent variable.

Please find the screenshot attached showing the Regression screenshot after eliminating multicollinearity.

