

# PROJECT 04



# IT PROJECTS

1. Data Preprocessing
2. Exploratory Data Analysis (EDA)
3. Correlation Analysis
4. Data Visualization

By:  
Mehrdad Mansourdehghan



GitHub



YouTube Channel



LinkedIn Profile



Send me E-mail

Project Goal .....	2
Language, libraries, tools .....	2
Data .....	2

## **Data Preprocessing**

Handle Duplicate Rows .....	3
Check Data Type .....	3
Enrich Data .....	4

## **Exploratory Data Analysis (EDA)**

Qualitative Variables .....	5
Quantitative Variables .....	6

## **Correlation Analysis**

Covariance and Correlation .....	8
ANOVA test .....	9

## **Data Visualization**

Dashboard .....	16
-----------------	----

## 1. Project Goals

In this project, I'm going to preprocess and clean data, run exploratory data analysis (EDA), find descriptive statistical measurements, and find the correlation among variables of a dataset that contains 99 project's data. Also, I will create a dashboard to visualize the data.

## 2. Language, libraries, tools:

Language: -

Libraries: Pivot Table, Data Analysis Tools

IDE: -

Application: Microsoft Excel

## 3. Data

There is an excel sheet that contains 12 columns with 99 records that are information about different projects:

- **Project type:** in the company there are four different project type: "Cost Reduction", "Income Generation", "Working Capital Improvement", and "Process Improvement".
- **Project Manager:** name of each project manager.
- **Region:** the region where the project is running. "East", "West", "North" and "South".
- **Department:** the department of the company that project is running for it.
- **Project Cost:** the budget or cost of the project in dollars.
- **Project Benefit:** the revenue of benefit that is anticipated that project makes.
- **Complexity:** it shows how complex every project is. "High", "Medium" and "Low".
- **Status:** the last status of the project in terms of progress. "Completed", "In Progress", "Cancelled" and "On hold"
- **Phase:** the last phase of each project based on PMBOK. "Initiation", "Planning", "Executing", "Closing"
- **Date:** the start and finish date of project based on schedule.
- **year:** year of release



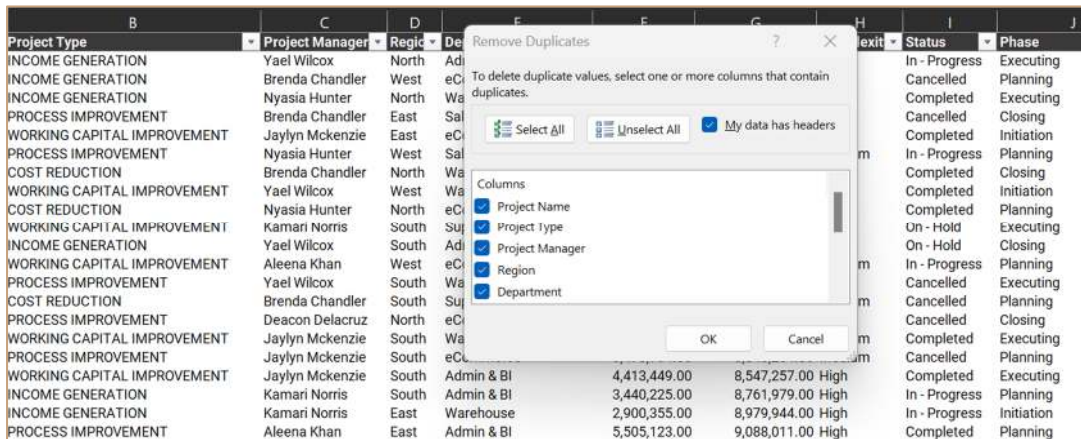
# Data Preprocessing

# Data Analysis Project: IT PROJECTS

Here you can follow all steps that were taken in this project.

## 1. Duplicate rows:

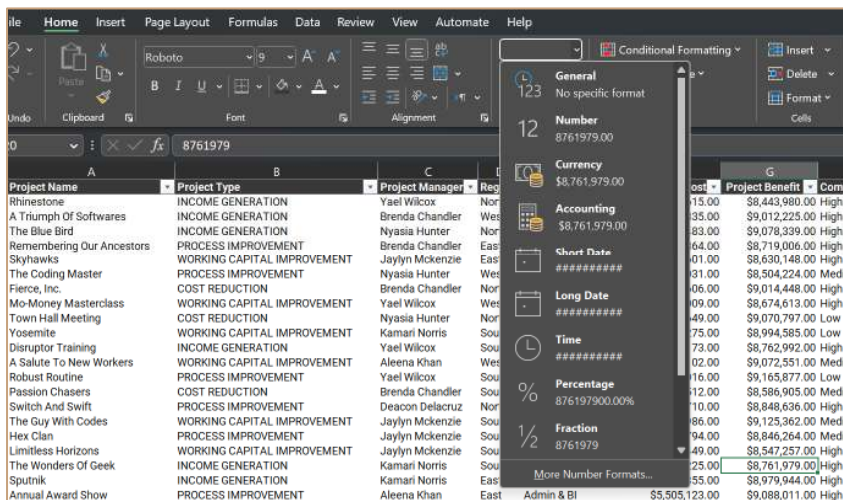
We checked duplicate values and hopefully we don't have two same rows that contain the same values.



Project Type	Project Manager	Region	Department	Status	Phase
INCOME GENERATION	Yael Wilcox	North	Admin & BI	In - Progress	Executing
INCOME GENERATION	Brenda Chandler	West	eC	Cancelled	Planning
INCOME GENERATION	Nyasia Hunter	North	Wa	Completed	Executing
PROCESS IMPROVEMENT	Brenda Chandler	East	Sal	Cancelled	Closing
WORKING CAPITAL IMPROVEMENT	Jaylyn McKenzie	East	eC	Completed	Initiation
PROCESS IMPROVEMENT	Nyasia Hunter	West	Sal	In - Progress	Planning
COST REDUCTION	Brenda Chandler	North	Wa	Completed	Closing
WORKING CAPITAL IMPROVEMENT	Yael Wilcox	West	Wa	Completed	Initiation
COST REDUCTION	Nyasia Hunter	North	eC	Completed	Planning
WORKING CAPITAL IMPROVEMENT	Kamari Norris	South	Suj	On - Hold	Executing
INCOME GENERATION	Yael Wilcox	South	Adj	On - Hold	Closing
WORKING CAPITAL IMPROVEMENT	Aleena Khan	West	eC	In - Progress	Planning
PROCESS IMPROVEMENT	Yael Wilcox	South	Wa	Cancelled	Executing
COST REDUCTION	Brenda Chandler	South	Suj	Cancelled	Planning
PROCESS IMPROVEMENT	Deacon Delacruz	North	eC	Cancelled	Closing
WORKING CAPITAL IMPROVEMENT	Jaylyn McKenzie	South	Wa	Completed	Executing
PROCESS IMPROVEMENT	Jaylyn McKenzie	South	eC	Cancelled	Planning
WORKING CAPITAL IMPROVEMENT	Jaylyn McKenzie	South	Admin & BI	Completed	Executing
INCOME GENERATION	Kamari Norris	South	Admin & BI	In - Progress	Planning
INCOME GENERATION	Kamari Norris	East	Warehouse	In - Progress	Initiation
PROCESS IMPROVEMENT	Aleena Khan	East	Admin & BI	Completed	Planning

## 2. Data type:

In the next, we should check the data type of each column to make sure they are presenting the value that they were supposed to do. "Project Name", "Project Type", "Project Manager", "Region", "Department", "Complexity", "Status" and "Phase" are text. "Project Cost" and "Project Benefit" are currency. Also, "Start Date", "End Date" are short dates.




Project Name	Project Type	Project Manager	Region	Department	Project Cost	Project Benefit	Complexity
Rhinestone	INCOME GENERATION	Yael Wilcox	North	Admin & BI	\$8,761,979.00	\$8,443,980.00	High
A Triumph Of Softwares	INCOME GENERATION	Brenda Chandler	West	eC	\$8,761,979.00	\$8,443,980.00	High
The Blue Bird	INCOME GENERATION	Nyasia Hunter	North	Wa	\$8,761,979.00	\$8,443,980.00	High
Remembering Our Ancestors	PROCESS IMPROVEMENT	Brenda Chandler	East	Sal	\$8,761,979.00	\$8,443,980.00	High
Skyhawks	WORKING CAPITAL IMPROVEMENT	Jaylyn McKenzie	East	eC	\$8,761,979.00	\$8,443,980.00	High
The Coding Master	PROCESS IMPROVEMENT	Nyasia Hunter	West	Sal	\$8,761,979.00	\$8,443,980.00	High
Fierce, Inc.	COST REDUCTION	Brenda Chandler	North	Wa	\$8,761,979.00	\$8,443,980.00	High
Mo-Money Masterclass	WORKING CAPITAL IMPROVEMENT	Yael Wilcox	West	Wa	\$8,761,979.00	\$8,443,980.00	High
Town Hall Meeting	COST REDUCTION	Nyasia Hunter	North	eC	\$8,761,979.00	\$8,443,980.00	High
Yosemite	WORKING CAPITAL IMPROVEMENT	Kamari Norris	South	Suj	\$8,761,979.00	\$8,443,980.00	High
Disruptor Training	INCOME GENERATION	Yael Wilcox	South	Adj	\$8,761,979.00	\$8,443,980.00	High
A Salute To New Workers	WORKING CAPITAL IMPROVEMENT	Aleena Khan	West	eC	\$8,761,979.00	\$8,443,980.00	High
Robust Routine	PROCESS IMPROVEMENT	Yael Wilcox	South	Wa	\$8,761,979.00	\$8,443,980.00	High
Passion Chasers	COST REDUCTION	Brenda Chandler	South	Suj	\$8,761,979.00	\$8,443,980.00	High
Switch And Swift	PROCESS IMPROVEMENT	Deacon Delacruz	North	eC	\$8,761,979.00	\$8,443,980.00	High
The Guy With Codes	WORKING CAPITAL IMPROVEMENT	Jaylyn McKenzie	South	Wa	\$8,761,979.00	\$8,443,980.00	High
Hex Clan	PROCESS IMPROVEMENT	Jaylyn McKenzie	South	eC	\$8,761,979.00	\$8,443,980.00	High
Limitless Horizons	WORKING CAPITAL IMPROVEMENT	Jaylyn McKenzie	South	Admin & BI	\$8,761,979.00	\$8,443,980.00	High
The Wonders Of Geek	INCOME GENERATION	Kamari Norris	South	Admin & BI	\$8,761,979.00	\$8,443,980.00	High
Sputnik	INCOME GENERATION	Kamari Norris	East	Warehouse	\$8,761,979.00	\$8,443,980.00	High
Annual Award Show	PROCESS IMPROVEMENT	Aleena Khan	East	Admin & BI	\$8,761,979.00	\$8,443,980.00	High

# Data Analysis Project: IT PROJECTS

## 3. Enrich data:

To enrich data and improve the quality of data, we can make a calculate column. For doing this, we use finish date and start date column to create a new column that shows project duration. Project duration is the difference between finish date and start date.

	C	D	E	F	G	H	I
	Project Manager	Department	Complexit	Regio	Start Date	End Date	Duration
N	Yael Wilcox	Admin & BI	High	North	2/1/2021	6/1/2021	120
N	Brenda Chandler	eCommerce	High	West	3/1/2021	6/1/2021	92
N	Nyasia Hunter	Warehouse	High	North	3/1/2021	6/1/2021	92
ENT	Brenda Chandler	Sales and Marketing	High	East	3/1/2021	6/1/2021	92
IMPROVEMENT	Jaylyn Mckenzie	eCommerce	High	East	3/1/2021	6/1/2021	92
ENT	Nyasia Hunter	Sales and Marketing	Medium	West	3/1/2021	6/1/2021	92
	Brenda Chandler	Warehouse	High	North	4/1/2021	7/1/2021	91
IMPROVEMENT	Yael Wilcox	Warehouse	High	West	4/1/2021	7/1/2021	91
	Nyasia Hunter	eCommerce	Low	North	5/1/2021	8/1/2021	92
IMPROVEMENT	Kamari Norris	Supply Chain	Low	South	5/1/2021	8/1/2021	92
N	Yael Wilcox	Admin & BI	High	South	5/1/2021	8/1/2021	92
IMPROVEMENT	Aleena Khan	eCommerce	Medium	West	5/1/2021	8/1/2021	92
ENT	Yael Wilcox	Warehouse	Low	South	5/1/2021	8/1/2021	92
	Brenda Chandler	Supply Chain	Medium	South	6/1/2021	9/1/2021	92
ENT	Deacon Delacruz	eCommerce	High	North	6/1/2021	9/1/2021	92
IMPROVEMENT	Jaylyn Mckenzie	Warehouse	Medium	South	7/1/2021	10/1/2021	92
ENT	Jaylyn Mckenzie	eCommerce	Medium	South	7/1/2021	10/1/2021	92
IMPROVEMENT	Jaylyn Mckenzie	Admin & BI	High	South	8/1/2021	11/1/2021	92
N	Kamari Norris	Admin & BI	High	South	9/1/2021	12/1/2021	91
N	Kamari Norris	Warehouse	High	East	10/1/2021	1/1/2022	92
ENT	Aleena Khan	Admin & BI	High	East	10/1/2021	1/1/2022	92



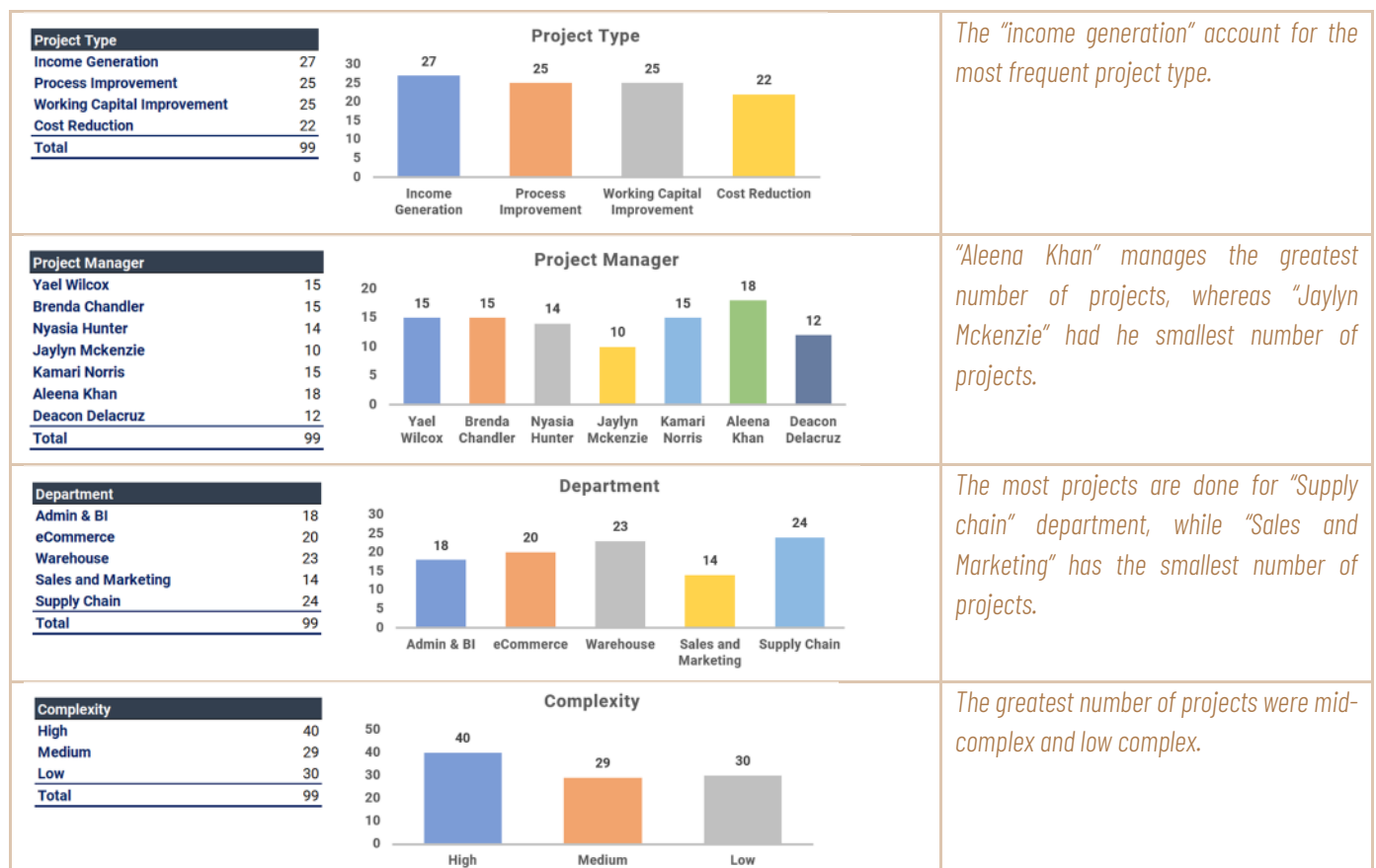
# Exploratory Data Analysis (EDA)

# Data Analysis Project: IT PROJECTS

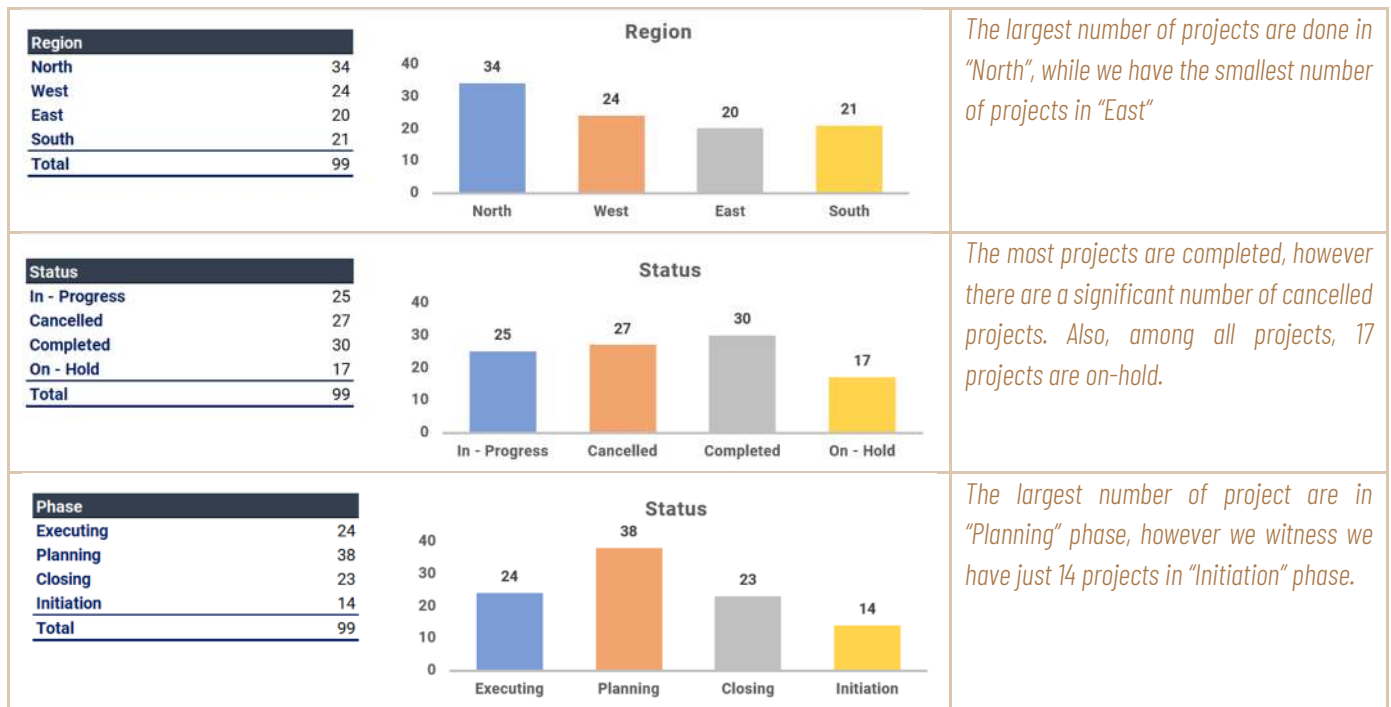
In this section we do descriptive statistical analysis to know data much more. By doing this, we can realize their distribution, central tendency measurement, the dispersion measurement and shape of the data.

## 1. Qualitative variables

Here we analyze qualitative variables, and we see each label in every single variable account for the highest frequency. Then, we can see the statistical interpretation for categorical variables. "Project Type", "Project Manager", "Department", "Complexity", "Region", "Status" and "Phase" are qualitative variables. So, I find the <distinct> values and then use <count if> function to find the frequency of each category.



# Data Analysis Project: IT PROJECTS



## 2. Quantitative variables

Now, let's analyze descriptive statics for quantitative variables. In this section we can find central tendency, dispersion, and shape measurements. Then we draw the distribution plot to compare with normal distribution. "Duration", "Project Cost" and "Project Benefit" are quantitative variables.

Duration			<p>The average duration of projects is 96.16 days.</p> <p>Half of the projects last less than 92 days</p> <p>The difference between the maximum and minimum duration is 30 days.</p> <p>25% of the projects take less than or equal to 91 days.</p> <p>75% of the projects take less than or equal to 92 days.</p> <p>50% of the projects take between 91 and 92 days.</p> <p>The distribution is right skewed.</p>
Mean	\$	96.16	
Median	\$	92.00	
Min	\$	90.00	
Max	\$	120.00	
Range	\$	30.00	
Q1	\$	91.00	
Q3	\$	92.00	
Var	\$	111.03	
St.d	\$	10.54	
Skew	\$	1.85	
Kurt	\$	1.49	

# Data Analysis Project: IT PROJECTS

Project Cost			<i>The average cost of projects is \$4,156,649. Half of the projects cost less than \$4,172,827. The difference between the maximum and minimum project cost is \$3,556,514. 25% of the projects cost less than or equal to \$3,251,948. 75% of the projects cost less than or equal to \$5,063,288. 50% of the projects cost between \$3,251,948 and \$5,063,288. The distribution is normal</i>
Mean	\$	4,156,649.36	
Median	\$	4,172,827.00	
Min	\$	2,418,301.00	
Max	\$	5,974,815.00	
Range	\$	3,556,514.00	
Q1	\$	3,251,948.00	
Q3	\$	5,063,288.50	
Var	\$	1,158,946,648,358.09	
St.d	\$	1,076,543.84	
Skew	\$	0.09	
Kurt	\$	(1.17)	
Project Benefit			<i>The average benefit of projects is \$8,828,178. Half of the projects could earn money less than \$8,846,243. The difference between the maximum and minimum gained benefit is \$743,299. 25% of the projects return benefit less than or equal to \$8,656,248. 75% of the projects return benefit less than or equal to \$9,019,233. 50% of the projects return benefit between \$8,656,248 and \$9,019,233. The distribution is normal</i>
Mean	\$	8,828,178.13	
Median	\$	8,846,243.00	
Min	\$	8,422,578.00	
Max	\$	9,165,877.00	
Range	\$	743,299.00	
Q1	\$	8,656,248.00	
Q3	\$	9,019,233.50	
Var	\$	46,829,775,907.71	
St.d	\$	216,401.89	
Skew	\$	(0.20)	
Kurt	\$	(1.27)	



# Correlation Analysis

## Data Analysis Project: IT PROJECTS

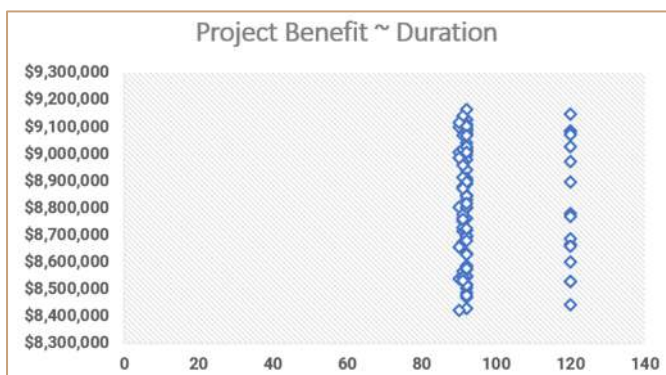
In this section, I'm going to see whether there is a relationship between variables with project benefits. Since the benefits is very crucial for each project, it makes sense that we see how many other factors are correlated to this variable. For doing this, we should consider two different ways. One way is correlation between numeric variables with benefit, and the second way is correlation between categorical variables with benefit. For the first one, we use Pearson correlation and for the second one, we use ANOVA test.

### 1. Covariance and Correlation

In this section we use correlation and covariance functions in excel. The covariance shows whether we have positive or negative relationships, and the correlation shows how strong that relationship is.

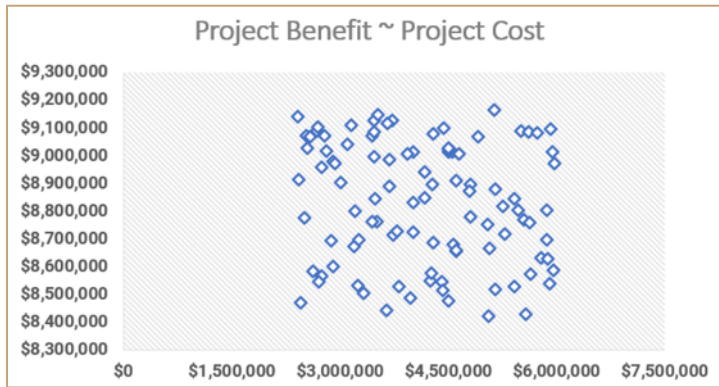
Project Benefit ~ Duration		Project Benefit ~ Project Cost	
Covariance	-90303.96021	Covariance	-36868465580
Correlation	-0.039601909	Correlation	-0.158256771

As we can see, there is a negative relationship between Duration of project and the money it pays back. The longer project leads to less benefit. However, this relationship is weak as if there is no relationship between them.



## Data Analysis Project: IT PROJECTS

Between project benefit and project cost we also have the negative relationship, and it shows if cost of project increase, the benefit witnesses reduction. Moreover, this relationship is moderate in terms of strength.



## 2. ANOVA Test

For categorical variables, we should see whether a particular categorical variable impacts project benefit or not (it is significant or not). For doing this we should use the ANOVA test for each categorical variable, and I do it with Pivot Table and Data Analysis tools in Microsoft Excel.

### Project Type

Average of Project Benefit	Column Labels			
Row Labels	Cost Reduction	Income Generation	Process Improvement	Working Capital Improvement
Grand Total	8844274.682	8812345.63	8889371.72	8769918.68

ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	1.90939E+11	3	63646229897	1.374686301	0.255254887	2.7004091
Within Groups	4.39838E+12	95	46298729992			
Total	4.58932E+12	98				

Since the p-value is more than 0.05, we accept the null hypothesis. It means "Project Type" is not significant variable to explain Project Benefits

# Data Analysis Project: IT PROJECTS

## Project Manager

Average of Project Benefit	Column Labels						
Row Labels	Aleena Khan	Brenda Chandler	Deacon Delacruz	Jaylyn Mckenzie	Kamari Norris	Nyasia Hunter	Yael Wilcox
Grand Total	8935128.111	8779876.133	8872598.75	8760820.8	8927122.067	8730203.929	8750007.2

ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	6.82828E+11	6	1.13805E+11	2.680163661	<b>0.019224479</b>	2.198779489
Within Groups	3.90649E+12	92	42461846251			
Total	4.58932E+12	98				

Since the p-value is less than 0.05, we reject the null hypothesis. It means "Project Manager" is a significant variable to explain Project Benefits

## Department

Average of Project Benefit	Column Labels			
Row Labels	Admin & BI	eCommerce	Sales and Marketing	Supply Chain Warehouse
Grand Total	8828263.333	8869463.6	8758809.5	8850856.417 8810771.13

ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	1.2077E+11	4	30192605744	0.63512917	<b>0.63869008</b>	2.468533034
Within Groups	4.46855E+12	94	47537740596			
Total	4.58932E+12	98				

Since the p-value is more than 0.05, we accept the null hypothesis. It means "Department" is not a significant variable to explain Project Benefits

## Complexity

Average of Project Benefit	Column Labels		
Row Labels	High	Low	Medium
Grand Total	8803544.4	8865352.233	8823699.724

ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	66311872361	2	33155936180	0.703728838	<b>0.49727105</b>	3.091191259
Within Groups	4.52301E+12	96	47114647569			
Total	4.58932E+12	98				

Since the p-value is more than 0.05, we accept the null hypothesis. It means "Complexity" is not a significant variable to explain Project Benefits

# Data Analysis Project: IT PROJECTS

## Region

Average of Project Benefit	Column Labels			
Row Labels	East	North	South	West
Grand Total	8801634.35	8849362.618	8796755.762	8847781.167

ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	59307417376	3	19769139125	0.414583623	0.7429143	2.700409063
Within Groups	4.53001E+12	95	47684322332			
Total	4.58932E+12	98				

Since the p-value is more than 0.05, we accept the null hypothesis. It means "Region" is not a significant variable to explain Project Benefits

## Status

Average of Project Benefit	Column Labels			
Row Labels	Cancelled	Completed	In - Progress	On - Hold
Grand Total	8865445.778	8859725.6	8734425.6	8851187.706

ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	2.96096E+11	3	98698630750	2.183993653	0.09497711	2.700409063
Within Groups	4.29322E+12	95	45191812071			
Total	4.58932E+12	98				


Since the p-value is more than 0.05, we accept the null hypothesis. It means "Status" is not a significant variable to explain Project Benefits

## Phase

Average of Project Benefit	Column Labels			
Row Labels	Closing	Executing	Initiation	Planning
Grand Total	8814566.87	8805784.833	8795323	8862664.184

ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	76601536572	3	25533845524	0.537528853	0.65770844	2.700409063
Within Groups	4.51272E+12	95	47502278972			
Total	4.58932E+12	98				

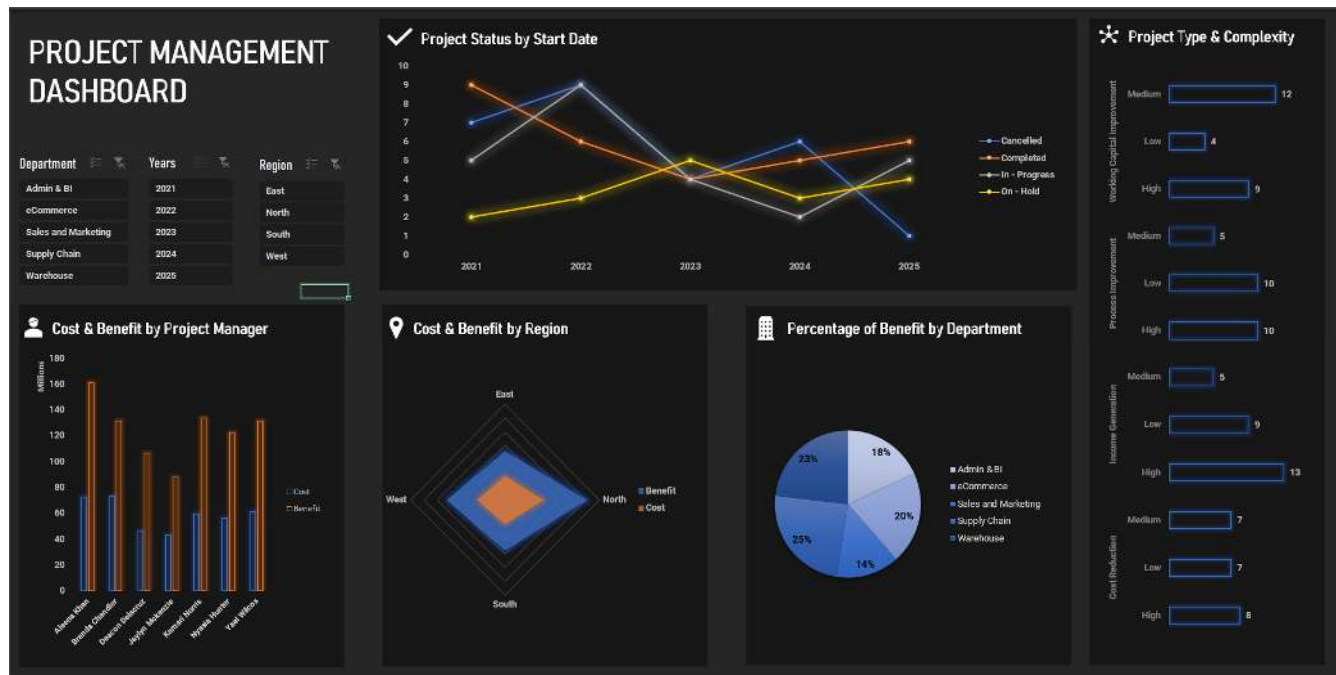
Since the p-value is more than 0.05, we accept the null hypothesis. It means "Phase" is not a significant variable to explain Project Benefits



# Data Visualization

## Data Analysis Project: IT PROJECTS

In this section we want to analyze data based on visualization. We use Microsoft Excel's charts and plots to make a dashboard. To make the dashboard, first I create five different pivot table sheets and in each of them I generate the needed report based on the data and after that, I choose the best and most appropriate charts.



In left side, we see the slicers of filter that we can use them for filtering data in dashboard. They are independent and we can choose some of them or all of them at the same time. At the bottom of these, we have a bar chart that shows a comparison between the cost and benefit of the project that is made by the project manager. In the top middle, we see a line chart that shows the number of project status in different years. Also, in the bottom, we different between cost and benefit in every region that gives us good view about money spend and gain in all regions as well as this we have a pie chart that shows the percentage of gained benefit for every department. Finally, we see another bar chart that shows a combination chart that shows number or project by its type and complexity.

— END —