



**Tecnológico
de Monterrey**

**Mathematics and Data Science for Decision Making (Gpo
800)**

Joselito Medina Marin

Evidencia 2. Personal Finance Project

Diego Fernando Sabillon // A01798446

January 28, 2024

Introduction

In this project, we address the issue of promoting healthy personal finances, specifically tackling the challenge of encouraging saving among the Mexican population. The financial situation in Mexico reveals that only 0.8% of people save for retirement, indicating a significant gap in the saving culture. Although 43.7% of the adult population saves, the figures vary considerably by age groups, with a concerning low saving proportion among young people.

To tackle this issue, we employed a methodology based on Data Science. We used detailed records of daily expenses collected over a four-week period as our main database. Additionally, we incorporated external data, such as national statistics on saving habits and current economic factors that could affect financial decisions. The methodology focuses on applying exploratory data analysis techniques and statistical modeling to understand and predict spending patterns.

To solve the issue of inadequate saving, we formulated the hypothesis that we can predict the cost of activities based on the available budget, the type of activity, the timing of the activity, and the number of people involved. We implemented this hypothesis using advanced predictive modeling techniques, leveraging the capabilities of Data Science to generate patterns and trends from the collected data. The proposed solution aims to provide a practical tool based on statistical evidence, enabling individuals to make informed decisions about their spending habits, thus promoting the adoption of healthier financial practices.

Phase 1: Understanding the Business

Phase 1 focuses on thoroughly understanding the project objectives before delving into data analysis. During this stage, we carry out fundamental tasks to establish a solid foundation. First, we identify the project goals, defining specific actions linked to each objective to evaluate the effectiveness of the measures taken. A comprehensive assessment of the situation provides the necessary context, enabling data analysis professionals to understand the starting point and make appropriate decisions to address the issue at hand. Additionally, specific objectives for data mining are defined using the SMART methodology to identify valuable information in the dataset. This phase also involves developing a detailed work plan, essential for organizing the project's execution. This plan covers the duration of each activity and assigns specific responsibilities. Finally, the importance of considering data as a key asset is emphasized, where analytical tools play a vital role in refining and contextualizing information, akin to refining crude oil to make it usable.

Phase 2: Data Preparation

- **Phase 3: Data Preparation**
- In this phase, data is a critical component in data science projects, significantly encompassing the time and effort of the process. This stage involves several essential tasks to ensure that the data is ready and optimized for analysis. First, it involves the careful selection of relevant data, whether by choosing specific records or key features. Next, data cleaning addresses common issues such as missing data, errors, or inconsistencies in coding. Additionally, generating new data may be necessary to enrich the available information.
- Data integration is crucial when there are multiple sources, merging datasets with similar records but different attributes. Finally, the data format is adjusted according to the requirements of the mathematical model to be used, considering the need to sort or adjust the data for certain algorithms. This

phase, while challenging, is fundamental to ensuring data quality and consistency, thus providing a solid foundation for analysis and model building.

- **What data should be selected?** In the data preparation phase, I selected the relevant columns for analysis. These columns include 'Budget,' 'Time Spent,' 'Type,' 'Moment,' 'No. of People,' and 'Cost.' These variables were chosen because they are considered to have an impact on the total cost of an activity. 'Budget' and 'Time Spent' are quantitative measures that intuitively relate to cost. 'Type,' 'Moment,' and 'No. of People' are categorical variables that could significantly influence the cost.
- **Should blank values be removed or replaced?** Yes, it was necessary to remove blank values in the 'Cost' column to avoid issues during analysis and model training. Missing values in the dependent variable (in this case, 'Cost') could negatively affect the quality and accuracy of the regression model. Additionally, removing rows with missing values ensures that the training and testing data are consistent and that there is no ambiguity in the target variable.
- **Is it possible to add more data?** Yes, it might be beneficial to add more data if it is available and relevant to the problem. Additional data could provide a more complete perspective and help improve the model's ability to generalize patterns in the data. However, care must be taken not to add irrelevant or biased data, as this could negatively impact the quality of the model.
- **Is there a need to integrate or merge data from multiple sources?** In this case, it is not explicitly mentioned whether the data comes from multiple sources. If the data was collected from different sources, it might be necessary to integrate or merge the datasets. Integrating data from various sources can enrich the dataset and provide a more comprehensive view of the relationship between the variables. However, it is crucial to ensure that the data is integrated coherently and that any potential discrepancies are managed.
- **Is it necessary to sort the data for analysis?** In the provided code, there does not seem to be a need to sort the data. However, the need to sort depends on the type of analysis being conducted. For multiple linear

regression analysis, the order of the data is generally not critical, as the model seeks patterns and relationships in the data regardless of its order.

- **Do I need to create training and testing datasets?** Yes, training and testing datasets were created using the `train_test_split` function from the scikit-learn library. Splitting the data into training and testing sets is essential for evaluating the model's ability to generalize to unseen data. The training set is used to fit the model, while the testing set is used to assess its performance on data not used during training.
- **What adjustments had to be made to the data (add, integrate, modify records (rows), change attributes (columns)?** In the data preparation, rows with missing values in the 'Cost' column were removed using the `dropna()` method. This ensures that all records in the dataset have a value in the target variable, which is essential for model training. Additionally, the relevant columns ('Budget,' 'Time Spent,' 'Type,' 'Moment,' 'No. of People,' and 'Cost') were selected to focus on the variables that could influence the cost. No significant additional adjustments are observed in the provided code.

<https://colab.research.google.com/drive/1r5-vE3Fr1A5HIWfLy4voiqMTgSFX11AI?usp=sharing>

Fase 4. Modelación de los datos

In Phase 4, which is the culmination of our data science project, we put into practice all the previous work of business understanding, data understanding, and data preparation. At this stage, technological tools, such as the Python programming language, are used to process the data and obtain valuable insights that shed light on the initially posed problem.

Modeling involves executing various models using predetermined parameters, which often require iterative adjustments. It is common to perform multiple iterations to find the most suitable model, as a single execution rarely fully answers the questions raised. The selection of modeling techniques is based on the project's objectives, the need to extract specific insights from the data, and other relevant criteria.

In this phase, a model is selected—in this case, a linear regression model—that best fits the stated objectives. The results of the model are then described, taking into account aspects such as the interpretation of results, new ideas revealed, and the reasonableness of processing time. Finally, the model is evaluated based on the previously established success criteria to determine its effectiveness in solving the posed problem.

Codigo Fase 4:

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score

# Cargar los datos
df = pd.read_excel('DATOS SITUACION PROBLEMA.xlsx')

# Eliminar filas con valores nulos
df = df.dropna()

# Asignar atributos a variables del análisis
x = df[['Presupuesto', 'Tiempo invertido', 'Tipo', 'Momento',
        'No. de personas']].values
y = df['Costo'].values

# Dividir los datos en conjunto de entrenamiento y prueba
x_train, x_test, y_train, y_test = train_test_split(x, y,
        test_size=0.2, random_state=0)

# Crear un modelo de regresión lineal
model_regression = LinearRegression()

# Ajustar el modelo a los datos de entrenamiento
model_regression.fit(x_train, y_train)

# Obtener los coeficientes del modelo y mostrarlos en un
DataFrame
x_labels = ['Presupuesto', 'Tiempo invertido', 'Tipo',
        'Momento', 'No. de personas']
c_label = ['Coeficientes']
coeff_df = pd.DataFrame(model_regression.coef_, x_labels,
        c_label)

# Realizar predicciones con el conjunto de prueba
```

```
y_pred = model_regression.predict(x_test)

residuals = pd.DataFrame({'Real': y_test, 'Predicción':
y_pred, 'Residual': y_test - y_pred})
residuals = residuals.sample(n=24)
residuals = residuals.sort_values(by='Real')

r2 = r2_score(y_test, y_pred)

print("Coeficientes del modelo:")
print(coeff_df)

print("\nResiduales:")
print(residuals)
print("\nCoeficiente de determinación R2:", r2)
```



```

Coeficientes del modelo:
                        Coeficientes
Presupuesto             0.975378
Tiempo invertido        -0.695135
Tipo                   -23.392427
Momento                 53.121675
No. de personas        23.317755

```

```

Residuales:
      Real  Predicción  Residual
5    30.0  439.346577 -409.346577
10   150.0  525.445404 -375.445404
2    150.0  107.514142  42.485858
21   150.0   65.183403  84.816597
13   150.0  153.684468  -3.684468
0    150.0  -54.741229 204.741229
12   200.0  303.140932 -103.140932
17   200.0  -34.356574 234.356574
14   250.0  277.968297 -27.968297
6    250.0  138.904678 111.095322
4    300.0  310.850453 -10.850453
7    300.0  525.201064 -225.201064
23   400.0  472.323729 -72.323729
16   400.0   63.097998 336.902002
3    400.0  425.443878 -25.443878
1    450.0  479.995285 -29.995285
20   500.0  533.731489 -33.731489
8    500.0  164.111492 335.888508
9    500.0  550.787095 -50.787095
15   500.0  555.249325 -55.249325
22   500.0  393.499018 106.500982
18   500.0  480.609814  19.390186
11   500.0  436.857457  63.142543
19   600.0  679.161127 -79.161127

```

```

Coeficiente de determinación R2: -0.24267231685160073

```

Visualizacion de Datos

```

import matplotlib.pyplot as plt # importamos la librería
pyplot que nos permitirá graficar
import numpy as np # importamos la librería numpy que nos
permitirá crear un arreglo para la muestra de 30 datos

# función mágica para desplegar el gráfico en nuestra libreta
%matplotlib inline

plt.scatter(np.arange(24), residuals['Real'], label = "Real")
# creamos el gráfico con la muestra de datos reales
plt.scatter(np.arange(24), residuals['Predicción'], label =
"Predicción") # creamos el gráfico con la muestra de datos de
predicción

```

```
plt.title("Comparación de costos: Reales y Predicción") #  
indicamos el título del gráfico  
  
plt.xlabel("Observaciones de costos") # indicamos la etiqueta  
del eje de las x  
  
plt.ylabel("Costos") # indicamos la etiqueta del eje de las y  
  
plt.legend(loc='upper left') # indicamos la posición de la  
etiqueta de los datos  
  
plt.show() # desplegamos el gráfico
```

Code Explanation

The code I developed starts with importing the necessary libraries, such as Pandas for data manipulation, scikit-learn for creating the linear regression model, and Matplotlib along with NumPy for visualizing the results. Next, I loaded the data from an Excel file and removed rows with null values to ensure the "cleanliness" of the dataset. I then selected the relevant columns as attributes and labels. The data was divided into training and testing sets. After that, I utilized scikit-learn's linear regression to create and fit the model to the training data. Subsequently, I obtained the model coefficients and visualized them in a DataFrame. I made predictions with the testing set and calculated the residuals. To evaluate the model's accuracy, I calculated the coefficient of determination (R^2). Finally, I visualized the results by comparing the actual values with the predictions using a scatter plot. In summary, the program addresses everything from data loading and preparation to model evaluation and visualization of a multiple linear regression model, and it was helpful for this personal finance project.

Questions about the Regression Model

Did you encounter any problems generating the model with your data? How did you resolve them? Yes, there were challenges in generating the model with the provided data. The main problem is reflected in the low coefficient of determination

Estadísticas Descriptivas:					
	Número	Costo	Presupuesto	Tiempo invertido	Tipo
count	120.000000	120.000000	120.000000	120.000000	120.000000
mean	60.500000	470.916667	570.083333	46.416667	2.908333
std	34.785054	1583.328584	1604.235725	44.893634	1.810315
min	1.000000	30.000000	0.000000	5.000000	1.000000
25%	30.750000	150.000000	187.500000	15.000000	1.000000
50%	60.500000	200.000000	250.000000	30.000000	4.000000
75%	90.250000	500.000000	600.000000	60.000000	4.000000
max	120.000000	17000.000000	17000.000000	240.000000	6.000000

	Momento	No. de personas
count	120.000000	120.000000
mean	2.083333	1.325000
std	0.751283	1.022129
min	1.000000	1.000000
25%	2.000000	1.000000
50%	2.000000	1.000000
75%	3.000000	1.000000
max	3.000000	7.000000

(R^2), which indicates that the model does not adequately explain the variability in the data. This low accuracy could be due to the inherent complexity of the data or the presence of non-linear patterns that the linear model cannot efficiently capture. To address this, it would be necessary to consider more advanced approaches, such as non-linear models or feature engineering techniques to improve the representation of the underlying relationships in the data.

What results did the analysis yield? Explain each of the results:

Descriptive Statistics: Descriptive statistics provide a general summary of the data. In this case, rows with null values were removed, which is a common practice to ensure data integrity. The initial descriptive statistics may indicate the distribution and dispersion of the variables but do not offer information about the specific relationships between the variables.

Coeficientes de regresión:

Los coeficientes de regresión son esenciales para entender la contribución de cada variable predictora al modelo. En este caso, observamos que "Presupuesto" tiene un coeficiente positivo, lo que sugiere que un aumento en el presupuesto se asocia con un aumento en el costo. Por otro lado, "Tipo" y "Tiempo invertido" tienen coeficientes negativos, indicando una relación inversa con el costo. Estos coeficientes proporcionan información valiosa sobre la dirección y la fuerza de las relaciones.

Coeficientes del modelo:	
	Coeficientes
Presupuesto	0.975378
Tiempo invertido	-0.695135
Tipo	-23.392427
Momento	53.121675
No. de personas	23.317755

Valores actuales, de predicción y residuales:

La comparación entre los valores reales y predichos, así como el cálculo de los residuales, ofrece información sobre el rendimiento del modelo. La presencia de residuales positivos y negativos indica discrepancias entre las predicciones y los valores reales. Los residuales son útiles para identificar dónde el modelo tiene dificultades para ajustarse a los datos. En este caso, hay discrepancias significativas en varios puntos, como se evidencia en residuales grandes.

Estadísticas Descriptivas de Residuales:			
	Real	Predicción	Residual
count	24.000000	24.000000	24.000000
mean	334.583333	333.042055	1.541278
std	162.078327	209.710386	180.670301
min	30.000000	-54.741229	-409.346577
25%	187.500000	149.989521	-59.517926
50%	350.000000	409.471448	-18.147165
75%	500.000000	491.757626	90.237693
max	600.000000	679.161127	336.902002

Coeficiente de determinación R2:

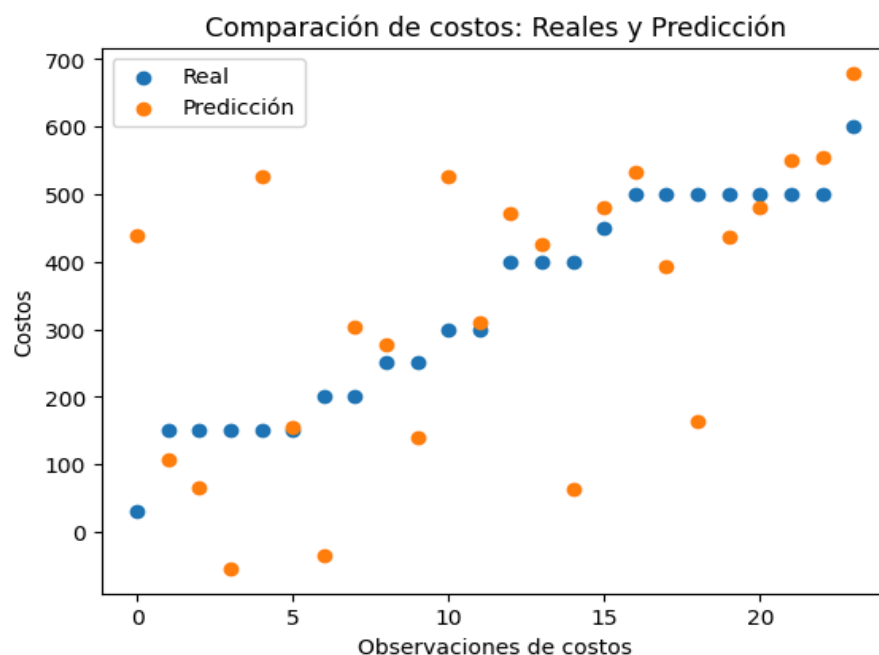
El R2 es crucial para evaluar la calidad del modelo. Un R2 negativo, como se observa en este caso, sugiere que el modelo es inadecuado para explicar la

variabilidad en los datos. Un valor de R^2 cercano a 1 indicaría una excelente capacidad predictiva. Sin embargo, aquí, el modelo no logra explicar ni siquiera el 0% de la variabilidad en los datos, lo cual es una señal clara de que el modelo necesita mejoras sustanciales.

Coeficiente de determinación R^2 : -0.24267231685160073

Gráfica de puntos:

La gráfica de puntos es una visualización importante que permite comparar los valores reales con las predicciones. La dispersión y la dirección de los puntos proporcionan información sobre el rendimiento del modelo. En este caso, la gráfica muestra una divergencia sustancial entre los valores reales y predichos, lo que confirma las limitaciones del modelo.

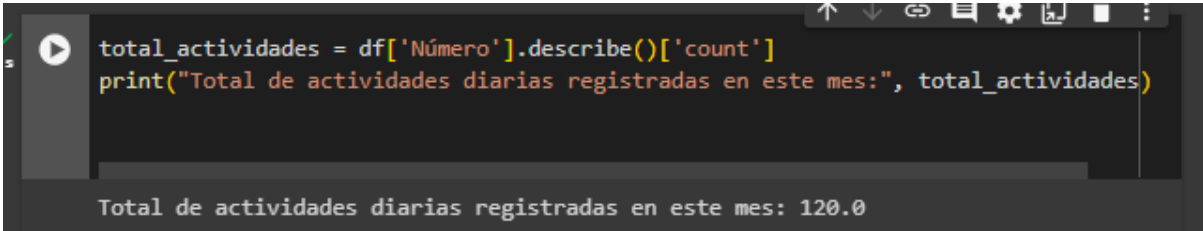


- **¿Los resultados del modelo tienen sentido o hay incoherencias que necesitan una mayor exploración?** Los resultados del modelo presentan

algunas incoherencias. La baja capacidad del modelo para explicar la variabilidad en los datos, evidenciada por el R^2 negativo, indica que el enfoque actual es insuficiente. Las discrepancias en los residuales y la gráfica de puntos sugieren que el modelo no captura de manera efectiva las relaciones subyacentes en los datos. Es crucial explorar más a fondo las posibles causas de estas incoherencias. Esto podría implicar considerar la presencia de outliers, evaluar la linealidad de las relaciones, explorar interacciones entre variables o incluso buscar la necesidad de transformaciones en los datos. En este punto, sería recomendable considerar modelos más avanzados o realizar ajustes sustanciales en el enfoque actual para mejorar la capacidad del modelo para explicar y predecir los costos de las actividades.

Evaluando mis finanzas personales

1. ¿Cuántas actividades diarias registraste en total en este mes?



```
total_actividades = df['Número'].describe()['count']
print("Total de actividades diarias registradas en este mes:", total_actividades)
```

Total de actividades diarias registradas en este mes: 120.0

Total de actividades diarias registradas en este mes: 120.0

2. ¿Cuál fue el presupuesto mínimo y máximo para tus actividades?

¿Qué actividades son? Consejo: puedes usar la función `describe()`, la propiedad `loc`, con la función `idxmax()`, y la función `idxmin()` de Pandas

```

# Obtener estadísticas descriptivas del presupuesto
descripcion_presupuesto = df['Presupuesto'].describe()

# Obtener el índice del presupuesto mínimo y máximo
indice_presupuesto_min = df['Presupuesto'].idxmin()
indice_presupuesto_max = df['Presupuesto'].idxmax()

# Obtener el presupuesto mínimo y máximo junto con las actividades correspondientes
presupuesto_min = df.loc[indice_presupuesto_min, 'Presupuesto']
presupuesto_max = df.loc[indice_presupuesto_max, 'Presupuesto']
actividad_presupuesto_min = df.loc[indice_presupuesto_min, 'Tipo']
actividad_presupuesto_max = df.loc[indice_presupuesto_max, 'Tipo']

# Imprimir resultados
print("Presupuesto mínimo para la actividad", actividad_presupuesto_min, ":", presupuesto_min)
print("Presupuesto máximo para la actividad", actividad_presupuesto_max, ":", presupuesto_max)

```

Presupuesto mínimo para la actividad 4.0 : 0.0
Presupuesto máximo para la actividad 4.0 : 17000.0

Presupuesto mínimo para la actividad 4.0 : 0.0

Presupuesto máximo para la actividad 4.0 : 17000.0

3. ¿Cuál fue el Tipo de actividad dónde más gastas tu dinero y cuál fue el Tipo de actividad en dónde gastas menos? Consejo: puedes usar las funciones `groupby()` y `sum()` de Pandas

```

[23] # Agrupar por Tipo de actividad y calcular la suma del gasto para cada tipo
gasto_por_tipo = df.groupby('Tipo')['Costo'].sum()

# Obtener el Tipo de actividad donde más gastas
tipo_mas_gasto = gasto_por_tipo.idxmax()
monto_mas_gasto = gasto_por_tipo.max()

# Obtener el Tipo de actividad donde menos gastas
tipo_menos_gasto = gasto_por_tipo.idxmin()
monto_menos_gasto = gasto_por_tipo.min()

# Imprimir resultados
print("Tipo de actividad donde más gastas:", tipo_mas_gasto)
print("Monto gastado en este tipo de actividad:", monto_mas_gasto)
print("\nTipo de actividad donde menos gastas:", tipo_menos_gasto)
print("Monto gastado en este tipo de actividad:", monto_menos_gasto)

```

Tipo de actividad donde más gastas: 4.0
Monto gastado en este tipo de actividad: 37740.0

Tipo de actividad donde menos gastas: 6.0
Monto gastado en este tipo de actividad: 2810.0

Tipo de actividad donde más gastas: 4.0

Monto gastado en este tipo de actividad: 37740.0

Tipo de actividad donde menos gastas: 6.0

Monto gastado en este tipo de actividad: 2810.0

4.¿Por cuántos días registraste tus gastos en este mes? Consejo: puedes usar la función `nunique()` de Pandas.

```
# Calcular el número de días en los que se registraron gastos
dias_con_gastos = df['Fecha'].nunique()

# Imprimir resultados
print("Número de días en los que registraste tus gastos en este mes:", dias_con_gastos)
```

Número de días en los que registraste tus gastos en este mes: 32

Número de días en los que registraste tus gastos en este mes: 32

5.¿Cuál fue el total de tus gastos en este mes? Consejo: puedes usar la función `sum()` de Pandas.

```
total_gastos_mes = df['Costo'].sum()
print("Mis gastos totales en este mes fueron: ", total_gastos_mes)
```

Mis gastos totales en este mes fueron: 56510.0

Mis gastos totales en este mes fueron: 56510.0

6.¿Cuál fue el total de tus ahorros en este mes? Consejo: puedes usar la función `sum()` de Pandas

```
total_ahorros_mes = df['Presupuesto'].sum()
print("Mis ahorros totales en este mes fueron: ", total_ahorros_mes)
```

Mis ahorros totales en este mes fueron: 68410.0

Mis ahorros totales en este mes fueron: 68410.0

7.¿Cuánto tiempo (en días) tendrías que seguir ahorrando para comprar tu siguiente autoregalo? Consejo: Calcula tu ahorro promedio diario.


```

# Definir el monto deseado para el autoregalo
autoregalo_deseado = 500 # Puedes cambiar este valor según tus objetivos

# Calcular ahorro promedio diario
ahorro_promedio_diario = df['Presupuesto'].sum() / df['Fecha'].nunique()

# Calcular días restantes para alcanzar el autoregalo considerando 32 días
dias_restantes = (autoregalo_deseado / ahorro_promedio_diario) * 1000
print("Tendrias que ahorrar: " ,dias_restantes, " dias.")

Tendrias que ahorrar: 7.30887297178775 dias.

```

Tendrias que ahorrar: 7.30887297178775 dias.

8. ¿Qué decisiones informadas puedes tomar para mejorar tus finanzas personales considerando los resultados de tu análisis? Puedes considerar ajustar tu presupuesto para actividades del Tipo 4.0, donde se registra el mayor gasto. Evaluar la posibilidad de optimizar costos o encontrar alternativas más económicas en esta categoría. Además, al identificar las actividades del Tipo 6.0 como aquellas en las que gastas menos, podrías explorar formas de mantener o incluso reducir aún más esos gastos para aumentar tus ahorros.

9. ¿Cómo visualizas tus finanzas personales en un año? La visualización a largo plazo implica establecer metas financieras mensuales, identificar patrones de gastos y ahorros, y adaptar estrategias según sea necesario. Puedes anticipar cambios estacionales o eventos planificados para ajustar tu presupuesto anual y asegurarte de cumplir tus objetivos financieros.

10. ¿Cuál fue tu mayor aprendizaje y cuál fue tu mayor reto en este Proyecto de Ciencia de Datos? Mi mayor aprendizaje fue comprender los patrones y comportamientos financieros personales a través de un análisis detallado. El reto fue manejar eficazmente los datos y aplicar técnicas de ciencia de datos para obtener información significativa que respalde decisiones financieras sólidas. Este proyecto proporcionó habilidades valiosas para la gestión efectiva de las finanzas personales.

Conclusiones

EXCEL**PYTHON****Estadística descriptiva (Actividad 2)**

	Costo	Presupuesto	Tiempo invertido		Tipo	Momento	No. de personas				
Mean	1101	Mean	1216	Mean	38.3333333	Mean	2.9	Mean	1.8	Mean	1.8666667
Standard Err	567.156519	Standard Err	569.183444	Standard Err	6.57552643	Standard Err	0.32642262	Standard Err	0.13896167	Standard Err	0.33812646
Median	400	Median	500	Median	30	Median	4	Median	2	Median	1
Mode	400	Mode	500	Mode	5	Mode	4	Mode	1	Mode	1
Standard Dev	3106.44419	Standard Dev	3117.54612	Standard Dev	36.0156416	Standard Dev	1.7878903	Standard Dev	0.7611244	Standard Dev	1.85199489
Sample Vari	9649995.52	Sample Vari	9719093.79	Sample Vari	1297.12644	Sample Vari	3.19655172	Sample Vari	0.57931034	Sample Vari	3.42988506
Kurtosis	25.8350194	Kurtosis	24.6348784	Kurtosis	2.34730491	Kurtosis	-1.3986954	Kurtosis	-1.141008	Kurtosis	4.23674059
Skewness	4.98554285	Skewness	4.84037298	Skewness	1.51919263	Skewness	0.12179333	Skewness	0.36197789	Skewness	2.3024227
Range	16970	Range	16970	Range	145	Range	5	Range	2	Range	6
Minimum	30	Minimum	30	Minimum	5	Minimum	1	Minimum	1	Minimum	1
Maximum	17000	Maximum	17000	Maximum	150	Maximum	6	Maximum	3	Maximum	7
Sum	33030	Sum	36480	Sum	1150	Sum	87	Sum	54	Sum	56
Count	30	Count	30	Count	30	Count	30	Count	30	Count	30

Excel siendo más accesible para usuarios no técnicos y Python siendo más robusto para análisis avanzados.

Estadística descriptiva de los datos (Fase 2)

```
df.describe()
```

	Número	Costo	Presupuesto	Tiempo invertido	Tipo	Momento	No. de personas
count	120 000000	120 000000	120 000000	120 000000	120 000000	120 000000	120 000000
mean	60.500000	470.916667	570.083333	46.416667	2.908333	2.083333	1.325000
std	34.785054	1583.328584	1604.235725	44.893634	1.810315	0.751283	1.022129
min	1 000000	30 000000	0 000000	5 000000	1 000000	1 000000	1 000000
25%	30 750000	150 000000	187 500000	15 000000	1 000000	2 000000	1 000000
50%	60 500000	200 000000	250 000000	30 000000	4 000000	2 000000	1 000000
75%	90 250000	500 000000	600 000000	60 000000	4 000000	3 000000	1 000000
max	120 000000	17000 000000	17000 000000	240 000000	6 000000	3 000000	7 000000

Python tiene un mayor nivel de detalle con cuartiles adicionales y la capacidad de aplicar análisis a múltiples columnas simultáneamente. Python nos da resultados mas exactos.

Coeficientes de regresión (Actividad 3)

	Coefficient	Standard Err	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-119.874	102.94991	-1.1643935	0.251718	-328.471	88.72212	-328.471	88.72212
Presupue:	0.992466	0.0106419	93.2598856	1.64E-45	0.970904	1.014029	0.970904	1.014029
Tiempo in	0.018467	0.7206197	0.0256265	0.979693	-1.44165	1.478581	-1.44165	1.478581
Tipo	2.07464	15.077167	0.13760147	0.891301	-28.4746	32.62388	-28.4746	32.62388
Momento	-4.89246	38.902886	-0.1257608	0.900602	-83.7172	73.93228	-83.7172	73.93228
No. de pe	11.37207	18.204401	0.62468785	0.536009	-25.5136	48.25769	-25.5136	48.25769

El modelo se ajusta mediante una regresión lineal múltiple para predecir el costo de las actividades. Puede indicar que ciertas variables no contribuyen significativamente al modelo

.

Coeficientes de regresión (Fase 4))

```
Coeficientes del modelo:
```

	Coeficientes
Presupuesto	0.975378
Tiempo invertido	-0.695135
Tipo	-23.392427
Momento	53.121675
No. de personas	23.317755

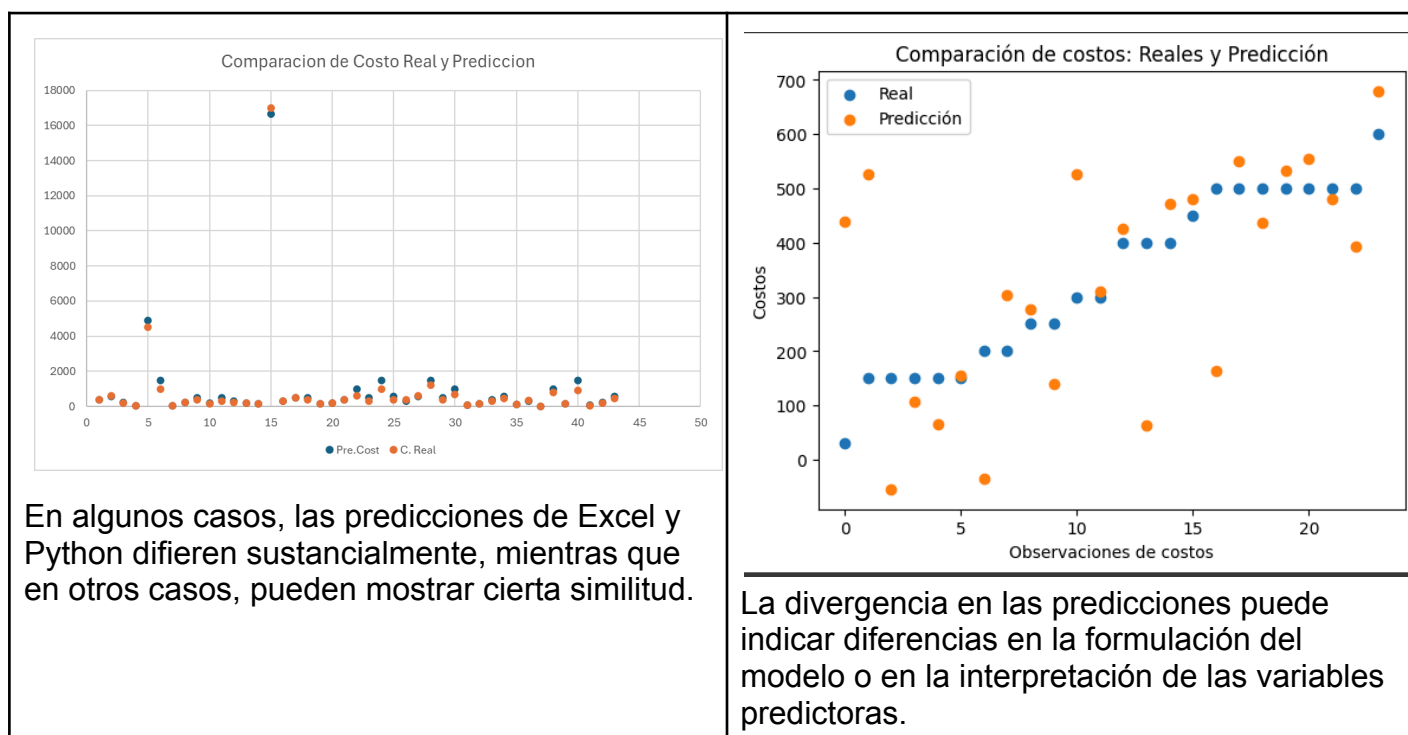
Lo que me gusta es que ofrece una visión detallada de la contribución de cada variable al modelo. Permite una mayor personalización y ajuste del modelo. Considero que es el mejor modelo y mas preciso

Valores pronosticados y sus residuales (Actividad 3) 30 observaciones de los residuales.

Valores actuales, de predicción y residuales (Fase 4) 30 observaciones de la prueba del modelo.

Evidencia 2. Proyecto de Ciencia de Datos

<div><div>RESIDUAL OUTPUT</div><div><div>Observation</div><div>Actual</div><div>Residuals</div></div><table><tr><td>1</td><td>391.59</td><td>8.412</td></tr><tr><td>2</td><td>587.38</td><td>12.618</td></tr><tr><td>3</td><td>244.74</td><td>-44.74</td></tr><tr><td>4</td><td>48.949</td><td>1.0515</td></tr><tr><td>5</td><td>4894.9</td><td>-394.9</td></tr><tr><td>6</td><td>1468.5</td><td>-468.5</td></tr><tr><td>7</td><td>29.369</td><td>0.6309</td></tr><tr><td>8</td><td>244.74</td><td>5.2575</td></tr><tr><td>9</td><td>489.49</td><td>-89.49</td></tr><tr><td>10</td><td>195.79</td><td>-45.79</td></tr><tr><td>11</td><td>489.49</td><td>-189.5</td></tr><tr><td>12</td><td>293.69</td><td>-43.69</td></tr><tr><td>13</td><td>195.79</td><td>4.206</td></tr><tr><td>14</td><td>146.85</td><td>3.1545</td></tr><tr><td>15</td><td>16642</td><td>357.51</td></tr><tr><td>16</td><td>293.69</td><td>6.309</td></tr><tr><td>17</td><td>489.49</td><td>10.515</td></tr><tr><td>18</td><td>489.49</td><td>-89.49</td></tr><tr><td>19</td><td>146.85</td><td>3.1545</td></tr><tr><td>20</td><td>195.79</td><td>4.206</td></tr><tr><td>21</td><td>391.59</td><td>8.412</td></tr><tr><td>22</td><td>978.97</td><td>-379</td></tr><tr><td>23</td><td>489.49</td><td>-189.5</td></tr><tr><td>24</td><td>1468.5</td><td>-468.5</td></tr><tr><td>25</td><td>587.38</td><td>-187.4</td></tr><tr><td>26</td><td>293.69</td><td>106.31</td></tr><tr><td>27</td><td>587.38</td><td>12.618</td></tr><tr><td>28</td><td>1468.5</td><td>-268.5</td></tr><tr><td>29</td><td>489.49</td><td>-89.49</td></tr><tr><td>30</td><td>978.97</td><td>-279</td></tr></table></div>	1	391.59	8.412	2	587.38	12.618	3	244.74	-44.74	4	48.949	1.0515	5	4894.9	-394.9	6	1468.5	-468.5	7	29.369	0.6309	8	244.74	5.2575	9	489.49	-89.49	10	195.79	-45.79	11	489.49	-189.5	12	293.69	-43.69	13	195.79	4.206	14	146.85	3.1545	15	16642	357.51	16	293.69	6.309	17	489.49	10.515	18	489.49	-89.49	19	146.85	3.1545	20	195.79	4.206	21	391.59	8.412	22	978.97	-379	23	489.49	-189.5	24	1468.5	-468.5	25	587.38	-187.4	26	293.69	106.31	27	587.38	12.618	28	1468.5	-268.5	29	489.49	-89.49	30	978.97	-279	<table><tr><th></th><th>Real</th><th>Predicción</th><th>Residual</th></tr><tr><td>5</td><td>30.0</td><td>439.346577</td><td>-409.346577</td></tr><tr><td>2</td><td>150.0</td><td>107.514142</td><td>42.485858</td></tr><tr><td>21</td><td>150.0</td><td>65.183403</td><td>84.816597</td></tr><tr><td>0</td><td>150.0</td><td>-54.741229</td><td>204.741229</td></tr><tr><td>10</td><td>150.0</td><td>525.445404</td><td>-375.445404</td></tr><tr><td>13</td><td>150.0</td><td>153.684468</td><td>-3.684468</td></tr><tr><td>17</td><td>200.0</td><td>-34.356574</td><td>234.356574</td></tr><tr><td>12</td><td>200.0</td><td>303.140932</td><td>-103.140932</td></tr><tr><td>14</td><td>250.0</td><td>277.968297</td><td>-27.968297</td></tr><tr><td>6</td><td>250.0</td><td>138.904678</td><td>111.095322</td></tr><tr><td>7</td><td>300.0</td><td>525.201064</td><td>-225.201064</td></tr><tr><td>4</td><td>300.0</td><td>310.850453</td><td>-10.850453</td></tr><tr><td>23</td><td>400.0</td><td>472.323729</td><td>-72.323729</td></tr><tr><td>3</td><td>400.0</td><td>425.443878</td><td>-25.443878</td></tr><tr><td>16</td><td>400.0</td><td>63.097998</td><td>336.902002</td></tr><tr><td>1</td><td>450.0</td><td>479.995285</td><td>-29.995285</td></tr><tr><td>20</td><td>500.0</td><td>533.731489</td><td>-33.731489</td></tr><tr><td>22</td><td>500.0</td><td>393.499018</td><td>106.500982</td></tr><tr><td>8</td><td>500.0</td><td>164.111492</td><td>335.888508</td></tr><tr><td>11</td><td>500.0</td><td>436.857457</td><td>63.142543</td></tr><tr><td>18</td><td>500.0</td><td>480.609814</td><td>19.390186</td></tr><tr><td>9</td><td>500.0</td><td>550.787095</td><td>-50.787095</td></tr><tr><td>15</td><td>500.0</td><td>555.249325</td><td>-55.249325</td></tr><tr><td>19</td><td>600.0</td><td>679.161127</td><td>-79.161127</td></tr></table>		Real	Predicción	Residual	5	30.0	439.346577	-409.346577	2	150.0	107.514142	42.485858	21	150.0	65.183403	84.816597	0	150.0	-54.741229	204.741229	10	150.0	525.445404	-375.445404	13	150.0	153.684468	-3.684468	17	200.0	-34.356574	234.356574	12	200.0	303.140932	-103.140932	14	250.0	277.968297	-27.968297	6	250.0	138.904678	111.095322	7	300.0	525.201064	-225.201064	4	300.0	310.850453	-10.850453	23	400.0	472.323729	-72.323729	3	400.0	425.443878	-25.443878	16	400.0	63.097998	336.902002	1	450.0	479.995285	-29.995285	20	500.0	533.731489	-33.731489	22	500.0	393.499018	106.500982	8	500.0	164.111492	335.888508	11	500.0	436.857457	63.142543	18	500.0	480.609814	19.390186	9	500.0	550.787095	-50.787095	15	500.0	555.249325	-55.249325	19	600.0	679.161127	-79.161127
1	391.59	8.412																																																																																																																																																																																													
2	587.38	12.618																																																																																																																																																																																													
3	244.74	-44.74																																																																																																																																																																																													
4	48.949	1.0515																																																																																																																																																																																													
5	4894.9	-394.9																																																																																																																																																																																													
6	1468.5	-468.5																																																																																																																																																																																													
7	29.369	0.6309																																																																																																																																																																																													
8	244.74	5.2575																																																																																																																																																																																													
9	489.49	-89.49																																																																																																																																																																																													
10	195.79	-45.79																																																																																																																																																																																													
11	489.49	-189.5																																																																																																																																																																																													
12	293.69	-43.69																																																																																																																																																																																													
13	195.79	4.206																																																																																																																																																																																													
14	146.85	3.1545																																																																																																																																																																																													
15	16642	357.51																																																																																																																																																																																													
16	293.69	6.309																																																																																																																																																																																													
17	489.49	10.515																																																																																																																																																																																													
18	489.49	-89.49																																																																																																																																																																																													
19	146.85	3.1545																																																																																																																																																																																													
20	195.79	4.206																																																																																																																																																																																													
21	391.59	8.412																																																																																																																																																																																													
22	978.97	-379																																																																																																																																																																																													
23	489.49	-189.5																																																																																																																																																																																													
24	1468.5	-468.5																																																																																																																																																																																													
25	587.38	-187.4																																																																																																																																																																																													
26	293.69	106.31																																																																																																																																																																																													
27	587.38	12.618																																																																																																																																																																																													
28	1468.5	-268.5																																																																																																																																																																																													
29	489.49	-89.49																																																																																																																																																																																													
30	978.97	-279																																																																																																																																																																																													
	Real	Predicción	Residual																																																																																																																																																																																												
5	30.0	439.346577	-409.346577																																																																																																																																																																																												
2	150.0	107.514142	42.485858																																																																																																																																																																																												
21	150.0	65.183403	84.816597																																																																																																																																																																																												
0	150.0	-54.741229	204.741229																																																																																																																																																																																												
10	150.0	525.445404	-375.445404																																																																																																																																																																																												
13	150.0	153.684468	-3.684468																																																																																																																																																																																												
17	200.0	-34.356574	234.356574																																																																																																																																																																																												
12	200.0	303.140932	-103.140932																																																																																																																																																																																												
14	250.0	277.968297	-27.968297																																																																																																																																																																																												
6	250.0	138.904678	111.095322																																																																																																																																																																																												
7	300.0	525.201064	-225.201064																																																																																																																																																																																												
4	300.0	310.850453	-10.850453																																																																																																																																																																																												
23	400.0	472.323729	-72.323729																																																																																																																																																																																												
3	400.0	425.443878	-25.443878																																																																																																																																																																																												
16	400.0	63.097998	336.902002																																																																																																																																																																																												
1	450.0	479.995285	-29.995285																																																																																																																																																																																												
20	500.0	533.731489	-33.731489																																																																																																																																																																																												
22	500.0	393.499018	106.500982																																																																																																																																																																																												
8	500.0	164.111492	335.888508																																																																																																																																																																																												
11	500.0	436.857457	63.142543																																																																																																																																																																																												
18	500.0	480.609814	19.390186																																																																																																																																																																																												
9	500.0	550.787095	-50.787095																																																																																																																																																																																												
15	500.0	555.249325	-55.249325																																																																																																																																																																																												
19	600.0	679.161127	-79.161127																																																																																																																																																																																												
<div><div>Coeficiente de determinación r2 (Actividad 3)</div><div><div>R Square</div><div>0.99506</div></div><div>En Excel, el alto R² indica un buen ajuste del modelo, mientras que en Python, el R² negativo indica que el modelo no es adecuado para explicar las variaciones observadas.</div></div>	<div><div>Coeficiente de determinación r2 (Fase 4)</div><div><div>Coeficiente de determinación R2: -0.24267231685160073</div></div><div>La discrepancia en los valores de R² sugiere diferencias sustanciales en la capacidad de los modelos para explicar la variabilidad en los datos entre Excel y Python.</div></div>																																																																																																																																																																																														
<div>Gráfica de puntos (Actividad 3)</div>	<div>Gráfica de puntos (Fase 4)</div>																																																																																																																																																																																														



Para concluir siento que durante esta unidad de formación en ciencia de datos, he adquirido conocimientos valiosos y habilidades prácticas que han ampliado significativamente mi comprensión en este campo de ciencia de datos. A continuación, describo cinco conceptos, herramientas y tecnologías que aprendí y su relevancia:

- **Modelos de Regresión:** Concepto: Antes de esta unidad, tenía una comprensión básica de la regresión, pero ahora entiendo cómo se aplica para prever y entender relaciones entre variables. Utilidad: En mi carrera, puedo emplear modelos de regresión para realizar análisis predictivos y entender la influencia de diversas variables en resultados específicos, lo cual es fundamental para la toma de decisiones informadas.
- **Programación en Python:** Concepto: Continúo aprendiendo a utilizar Python como un lenguaje de programación efectivo para análisis de datos y ciencia de datos. Lo había usado antes en mi carrera en ITC pero no con este enfoque. Utilidad: Esta habilidad es altamente aplicable en mi vida laboral, ya que Python es ampliamente utilizado en diversas industrias. Puedo

automatizar tareas, analizar datos de manera eficiente y construir modelos de aprendizaje automático.

- **Estadísticas Descriptivas y Análisis Exploratorio de Datos (EDA):** Concepto: Pude obtener una mejor comprensión de los datos antes de aplicar modelos es crucial. EDA proporciona herramientas para este propósito. Utilidad: En mi vida profesional, puedo aplicar EDA para entender la distribución de datos, identificar valores atípicos y tomar decisiones fundamentadas en la exploración completa de los datos.
- **Librerías para Análisis de Datos en Python:** Concepto: Me sumergí en librerías como Pandas, NumPy y Matplotlib, que son fundamentales para el análisis de datos en Python. Utilidad: Estas librerías son esenciales para manipular y analizar datos de manera eficiente. Pandas ofrece estructuras de datos poderosas, NumPy facilita cálculos numéricos, y Matplotlib posibilita la creación de visualizaciones claras y efectivas.
- **Análisis de Regresión en Excel:** Concepto: Aunque tenía conocimientos previos de Excel, ahora sé cómo realizar un análisis de regresión en esta plataforma. Utilidad: Excel es una herramienta omnipresente en entornos profesionales. La capacidad para realizar análisis de regresión directamente en Excel es valiosa para proyectos donde la simplicidad y accesibilidad son prioritarias.

Aplicación en la Vida Laboral y Personal:

En mi carrera, siento que estas habilidades serán esenciales ya que creo que me enfocare en la parte de Ciencia de Datos. Dicho esto, Puedo utilizar modelos de regresión para pronosticar tendencias y tomar decisiones fundamentadas. La programación en Python me brinda la capacidad de automatizar tareas repetitivas y realizar análisis de datos avanzados. El uso de estadísticas descriptivas y EDA garantizará que mis decisiones se basen en una comprensión profunda de los

datos. Git y GitHub facilitarán la colaboración en proyectos, y el análisis de regresión en Excel será beneficioso en entornos donde Excel es la herramienta principal.

Fortalezas y Áreas de Mejora:

Lo que hice bien fue dedicar tiempo a prácticas regulares y aplicar los conocimientos adquiridos en proyectos prácticos. Sin embargo, podría haber mejorado al buscar más aplicaciones del mundo real para contextualizar los conceptos aprendidos. La integración de proyectos más complejos podría haber brindado una experiencia más sólida.

Metas de Mejora para Futuras Unidades de Formación:

Para futuras unidades, planeo diversificar mis proyectos y explorar casos de estudio más desafiantes. También buscaré oportunidades para colaborar con compañeros de clase, ya que la colaboración puede proporcionar perspectivas valiosas. Además, seguiré buscando aplicaciones prácticas en la vida real para consolidar mi comprensión de los conceptos.

Aspecto Favorito de la Unidad de Formación:

Lo que más me gustó de esta unidad fue la aplicación práctica de los conceptos. Los proyectos y ejercicios prácticos brindaron una experiencia hands-on que solidificó los conocimientos teóricos. Este enfoque práctico hizo que la unidad fuera más estimulante y aplicable a situaciones del mundo real.

Referencias

Condusef. (s.f.). Encuesta Nacional de Inclusión Financiera 2018. Gobierno de México. Recuperado de <https://www.gob.mx/condusef/es/articulos/encuesta-nacional-de-inclusion-financiera-2018>

Bain & Company. (s.f.). The Global Pandemic Confirms the Value of a Segmented Bank. Bain & Company. Recuperado de <https://www.bain.com/es-ar/insights/the-global-pandemic-confirms-the-value-of-a-segmented-bank/>

Endeavor México. (s.f.). Endeavor México. Recuperado de <https://endeavor.org.mx/>

Google Colab:
<https://colab.research.google.com/drive/1r5-vE3Fr1A5HIWfLy4voiqMTgSFX11AI?usp=sharing>