**Logotipo

Descripción generada automáticamente**

**Análisis y Proyección de Costos en Proyectos de Vivienda VIS y NO VIS Desarrollados por la Constructora Vértice Ingeniería y Construcción en el Área Metropolitana del Valle de Aburrá durante el Período 2016-2023**

Eladio Yovera Yovera

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Asesor  
David Villanueva Valdes, MSc Data Science

Universidad de Antioquia  
Facultad de Ingeniería

Especialización en Analítica y Ciencia de Datos

Medellín, Antioquia, Colombia

2024

|  |  |
| --- | --- |
| **Cita** | (Yovera Yovera. 2024) |
| **Referencia**  **Estilo APA 7 (2020)** | Yovera Yovera, E. (2024). *Análisis y Proyección de Costos en Proyectos de Vivienda VIS y NO VIS Desarrollados por la Constructora Vértice Ingeniería y Construcción en el Área Metropolitana del Valle de Aburrá durante el Periodo 2016-2023.* Universidad de Antioquia, Medellín, Colombia. |

**** 

Especialización en Analítica y Ciencia de Datos, CohorteV.

Centro de Investigación Ambientales y de Ingeniería (CIA).

|  |  |
| --- | --- |
|  | Diagrama  Descripción generada automáticamente con confianza media |

Centro de Documentación Ingeniería (CENDOI)

**Repositorio Institucional:** http://bibliotecadigital.udea.edu.co

Universidad de Antioquia - www.udea.edu.co

Rector: John Jairo Arboleda Céspedes.

Decano: Julio Cesar Saldarriaga Molina

Jefe departamento: Diego José Luis Botia Valderrama

El contenido de esta obra corresponde al derecho de expresión de los autores y no compromete el pensamiento institucional de la Universidad de Antioquia ni desata su responsabilidad frente a terceros. Los autores asumen la responsabilidad por los derechos de autor y conexos.

**Dedicatoria**

A mi familia y amigos, por su amor y aliento. A la Constructora Vértice Ingeniería y Construcción y a su Gerente Comercial y Desarrollo Mateo Ramirez que permitió utilizar sus datos, los profesores por sus comentarios y sugerencias constructivas. Y a todos aquellos que han creído en mí a lo largo del camino. Este trabajo es un testimonio de su apoyo y dedicación.

**Agradecimientos**

En primer lugar, agradecer a Dios, por las fuerzas y bendiciones que ha derramado sobre mi vida. También quiero agradecer a mi familia y amigos por su amor, aliento y comprensión durante mis estudios. Sus palabras de apoyo me han motivado a seguir adelante en los momentos difíciles.

Agradezco a los profesores, por sus valiosos comentarios y sugerencias que han mejorado significativamente la calidad de este trabajo.

Este proyecto no habría sido posible sin el apoyo de estas personas. A todos ellos, les expreso mi más sincero agradecimiento.

**Tabla de contenido**

Contents

[Resumen 8](#_Toc167948728)

[Abstract 9](#_Toc167948729)

[1. Descripción del problema 10](#_Toc167948730)

[1.1. Problema de negocio 10](#_Toc167948731)

[1.2. Aproximación desde la analítica de datos 10](#_Toc167948732)

[1.3. Origen de los datos 11](#_Toc167948733)

[1.4. Métricas de desempeño 11](#_Toc167948734)

[2. Objetivos 12](#_Toc167948735)

[2.1. Objetivo general 12](#_Toc167948736)

[2.2. Objetivos específicos 12](#_Toc167948737)

[3. Datos 13](#_Toc167948738)

[3.1. Datos originales 13](#_Toc167948739)

[3.2. Datasets 16](#_Toc167948740)

[3.3. Analítica descriptiva 16](#_Toc167948741)

[18](#_Toc167948742)

**Lista de tablas**

[**Tabla 1**](#_heading=h.2u6wntf) Proyectos 13

**Tabla 2** Insumos 14

[**Tabla 3** Tipos de costo 1](#_heading=h.3tbugp1)5

**Tabla 4** Control de Proyectos 15

**Tabla 5** Dataset Final 16

**Lista de figuras**

[**Figura 1** Lista de variables categóricas](#_heading=h.2r0uhxc) 18

[**Figura 2** Lista de variables numéricas](#_heading=h.1664s55) 18

**Figura 3** Frecuencia de instancias de variables categóricas 19

**Figura 4** Relación entre variables numéricas 20

**Figura 5** Histograma y caja de bigotes del Costo del Proyecto 21

**Figura 6** Caja de bigotes del Costo del Proyecto después de eliminar outliers 22

**Figura 7** Histograma y caja de bigotes del Costo del Proyecto después de las transformaciones 23

**Figura 8** Mapa de calor de la matriz de correlaciones 24

**Figura 9** Valores de las correlaciones más fuertes 25

**Siglas, acrónimos y abreviaturas**

**APA** American Psychological Association

**CD** Ciencia de Datos

**CSV** Tipo de formato de archivo con valores separados por comas

**Esp.** Especialista

**INT** Tipo de dato numérico entero

**IVA** Impuesto al Valor Agregado

**MAE** Error medio absoluto

**NO VIS Vivienda de no Interés social**

**Párr.** Párrafo

**R²** Coeficiente de Determinación

**RMSE** Raíz cuadrada del error medio cuadrático

**VIS Vivienda de Interés Social**

**UdeA** Universidad de Antioquia

**RLM** Regresión Lineal Múltiple

# Resumen

El proyecto tiene como objetivo proyectar el costo de proyectos de vivienda VIS y NO VIS utilizando datos históricos de la Constructora Vértice Ingeniería y Construcción, abarcando el periodo de 2016 a 2023. Se utilizará un modelo de Machine Learning de Regresión Lineal Múltiple para realizar estas proyecciones. Inicialmente, se llevó a cabo un análisis descriptivo de los datos, donde se identificó que las variables categóricas y numéricas no presentaban una relación significativa, con valores de correlación menores a 0.09. Este hallazgo conllevará a una reestructuración de las variables, cambiando el enfoque del análisis de proyectos completos a análisis por insumos específicos.

Durante el desarrollo del proyecto, se emplearon técnicas avanzadas de preprocesamiento de datos, como la normalización y la codificación de variables categóricas, para mejorar la calidad de los datos utilizados en el proceso. A pesar de estos esfuerzos, se encontraron varios obstáculos, como la falta de datos consistentes y la necesidad de manejar datos faltantes y outliers. Sin embargo, mediante transformaciones y ajustes continuos de las variables, se lograron mejoras significativas en la precisión de las correlaciones.

Los resultados obtenidos indican que, con la reestructuración de las variables y el enfoque por insumos, el modelo de Regresión Lineal Múltiple podría proporcionar estimaciones más precisas y confiables. Esto permitirá a la constructora optimizar la planificación financiera, reducir sobrecostos y mejorar la eficiencia en la ejecución de sus proyectos de vivienda.

*Palabras clave*: Regresión, Outliers, Machine, Proyección, Vivienda, VIS, NO VIS.

Repositorio GitHub: <https://github.com/eladioyovera/specialization_udea>

# Abstract

The project aims to project the cost of VIS and NON-VIS housing projects using historical data from Constructor Vértice Ingeniería and Construcción, covering the period from 2016 to 2023. A Multiple Linear Regression Machine Learning model will be used to make these projections. Initially, a descriptive analysis of the data was carried out, where it was identified that the categorical and numerical variables did not present a significant relationship, with correlation values ​​less than 0.09. This finding will lead to a restructuring of the variables, changing the focus from the analysis of complete projects to analysis by specific inputs.

During the development of the project, advanced data preprocessing techniques, such as normalization and categorical variable coding, were used to improve the quality of the data used in the process. Despite these efforts, several obstacles were encountered, such as the lack of consistent data and the need to handle missing data and outliers. However, through continuous transformations and adjustments of the variables, significant improvements in the precision of the correlations were achieved.

The results obtained indicate that, with the restructuring of the variables and the input approach, the Multiple Linear Regression model could provide more precise and reliable estimates. This will allow the construction company to optimize financial planning, reduce cost overruns and improve efficiency in the execution of its housing projects.

*Keywords***:** Regresión, Outliers, Machine, Proyección, Vivienda, VIS, NO VIS.

# Descripción del problema

## Problema de negocio

La industria de la construcción en Colombia enfrenta un desafío crítico en la estimación precisa de costos de proyectos, lo que genera repercusiones negativas en el sector.

De acuerdo con un estudio de la Constructora Contex, en los últimos cinco años, la industria ha experimentado un panorama complejo debido a factores como la planeación ineficiente de proyectos, altas tasas de interés, disminución de subsidios y aumento en el precio de materiales. Esto ha resultado en la liquidación de aproximadamente 150 constructoras y la reestructuración de 366 más por parte de la Superintendencia de Sociedades - SuperSociedades. (Contex, 2023). Por esta razón, se entiende que la estimación de costos en la construcción es un proceso complejo y sujeto a una gran cantidad de incertidumbres, lo que genera grandes retos, tales como; Sobrecostos, dónde el costo real de los proyectos supera significativamente lo presupuestado; Retrasos, dónde la imprecisión en la estimación de costos conduce a retrasos en la ejecución de los proyectos. Mediante el análisis de grandes conjuntos de datos históricos y la aplicación de técnicas de aprendizaje automático, se desarrollará modelos predictivos más precisos y confiables para estimar costos de manera eficiente. Asimismo, los datos provendrán internamente de la Constructora desde el año 2016 hasta el año 2023. Las métricas de desempeño podrían incluir la precisión del modelo y el valor del costo de los proyectos futuros con características similares a los actuales.

## Aproximación desde la analítica de datos

Para abordar este problema, se desarrollará un modelo de Regresión Lineal Múltiple (RLM) debido a su robustez y capacidad para explicar la relación entre una variable dependiente (costo del proyecto de vivienda) y múltiples variables independientes (características del proyecto). Por otro lado, de acuerdo con la literatura y aplicaciones reales, la implementación de esta técnica es relativamente sencilla y computacionalmente eficiente, lo que facilita su aplicación en entornos prácticos. Además, El RLM puede manejar tanto variables categóricas como numéricas, permitiendo modelar una amplia gama de factores que influyen en el costo de los proyectos de vivienda.

## Origen de los datos

Tenemos datos de proyectos de vivienda VIS y NOS VIS, obtenidos de la Constructora Vértice Ingeniería y Construcción, los datos van a incluir información sobre las características del proyecto, los costos directos, costos indirectos, tipo de proyecto, insumos y otras variables relevantes para predecir el precio del costo del proyecto. Los datos comprenden desde el año 2016 hasta el año 2023, durante este período se desarrollaron 35 proyectos de vivienda.

## Métricas de desempeño

Dado que se ha seleccionado la técnica de Regresión Lineal Múltiple, las métricas de desempeño a evaluar incluirán el Error medio absoluto (MAE), el cual permitirá medir la diferencia promedio entre el costo real del proyecto y el costo predicho por el modelo, así como también, la Raíz cuadrada del error medio cuadrático (RMSE) el cual permitirá penalizar los errores grandes que se encuentren en el MAE. Otra de las métricas a considerar es el Coeficiente de Determinación (R²) el cual indicará la proporción de la variabilidad del costo total del proyecto que es explicada por el modelo. Adicionalmente, se utilizará la métrica de Precisión, Recall y F1-score, el primero indicará el porcentaje de proyectos en los que el costo predicho por el modelo está dentro de un rango de error predefinido, el segundo indicará el porcentaje de proyectos en los que el modelo identifica correctamente que el costo real está por encima del presupuesto y finalmente el tercero para medir la combinación de Precisión y Recall.

Con respecto a las métricas se enfatizará mucho en el valor de R², el cual se desea obtener un valor que se encuentre en un rango de 0.7 a 0.9, con base a ello se tomarán decisiones de viabilidad del proyecto.

# Objetivos

## Objetivo general

Desarrollar e implementar un sistema para estimar con mayor precisión y eficiencia los costos de proyectos de construcción de vivienda VIS y NO VIS que permita reducir los sobrecostos en proyectos de construcción y mejorar la planificación financiera y la toma de decisiones de la Constructora Vértice Ingeniería y Construcción basado en datos de los años 2016 al 2023.

## Objetivos específicos

* Identificar fuentes de información: Buscar dentro de la organización el repositorio de datos, en el cual se encuentre la información de los proyectos de vivienda.
* Recolectar y preprocesar los datos: Obtener y limpiar los datos de los proyectos de la constructora para su uso en el modelo de regresión lineal múltiple
* Desarrollar un modelo de regresión: Utilizar un modelo de regresión lineal múltiple, dado que, se tiene múltiples variables independientes, tales como, el tipo de costo, tipo de proyecto etc.
* Evaluar y optimizar el modelo: Evaluar el modelo utilizando un conjunto de datos de prueba y optimizarlo para mejorar su precisión.

# Datos

## Datos originales

Durante la identificación de fuentes de información, resulta que los datos de los proyectos se encuentran en una base de datos SQL Server On Premise, distribuidos en cuatro tablas, que al final eran parte de un modelo de datos tipo estrella (tres dimensiones y una tabla de hechos). La data comprende información desde el año 2016 hasta el año 2023. Para su uso se descargaron a través de archivos en formato CSV y en conjunto ocupan 500 MB de información. Los archivos contienen información de Nombre de Proyectos que incluyen 268 registros, Insumos que incluyen 110000 registros, Tipos de Costos que incluyen 3227 registros y finalmente la tabla de hechos Control Proyectos donde se encuentran los campos para relacionar las dimensiones y el valor del costo del proyecto, incluye un millón de registros.

El modo de acceso a la data es mediante una máquina virtual, para ello se debe tener las credenciales que otorga la Constructora Vértice Ingeniería y Construcción para poder acceder.

A continuación, se muestra la descripción de cada uno de los campos de las tablas mencionadas anteriormente.

**Tabla 1**  
*Proyectos*

|  |  |  |
| --- | --- | --- |
| **Nombre Columna** | **Tipo De Dato** | **Descripción** |
| SkIdProyecto | Int | Id de relación contra las otras tablas por medio del proyecto |
| Nombre Proyecto | varchar | Nombre del proyecto |
| Clase Proyecto | varchar | Clase a la cual pertenece el proyecto |
| Tipo | varchar | Tipo de proyecto que puede ser ADPRO, CBR o Ambas |
| Estado | varchar | Estado en el que se encuentra el proyecto los cuales pueden ser Presupuesto, Inactivo, Finalizado, En ejecución |
| Presupuesto Fijo | Char | Muestra si tiene presupuesto fijo o no |
| Propietario | varchar | Propietario del proyecto |
| Sucursal | smallint | Código de la sucursal asociada al proyecto |
| Sucursal Nombre | varchar | Nombre de la sucursal asociada al proyecto |
| MacroProyecto | varchar | Código del macro proyecto al cual pertenece el proyecto |
| MacroProyecto Descripcion | varchar | Descripción del macro proyecto al cual pertenece el proyecto |
| Centro Costo | varchar | Código del centro de costo al cual pertenece el proyecto |
| Centro Costo Descripcion | varchar | Descripción del centro de costo al cual pertenece el proyecto |
| VIS | Char | El proyecto es de tipo VIS (SI o NO) |
| Sucursal Administrativa | varchar | Descripción de la sucursal administrativa del proyecto |
| SkIdEmpresa | smallint | Id de relación contra las otras tablas por medio de la empresa |
| Empresa | nvarchar | Nombre de la empresa que se encuentra asociado el proyecto |

**Tabla 2**  
*Insumos*

|  |  |  |
| --- | --- | --- |
| **Nombre Columna** | **Tipo De Dato** | **Descripción** |
| SkIdEmpresa | smallint | Id del insumo, el cual permite relación con las otras tablas. |
| Empresa | nvarchar | Nombre de la empresa que se encuentra asociado al insumo. |
| SkIdInsumo | Int | Id del insumo, el cual permite relación con las otras tablas. |
| Insumo Descripcion | varchar | Descripción del insumo |
| Agrupacion | varchar | Id de la Agrupación asignada al insumo. |
| Agrupacion Descripcion | varchar | Descripción de la agrupación asignada al insumo. |
| Tipo | Char | Id del Tipo de Insumo asignado. |
| Tipo Descripcion | varchar | Descripción del tipo de insumo asignado. |
| Unidad | varchar | Id de la unidad de medida asignada al insumo. |
| Descripcion Unidad | varchar | Descripción de la unidad de medida asignada al insumo. |
| Estado | varchar | Estado actual del insumo. |
| Requiere Equipo | varchar | Configuración del insumo que permite ver el nivel de requerimiento del equipo. |
| Dias Reposicion | Int | Número de días de reposición configurados en el insumo. |
| SubAnalisis | Char | Configuración del insumo el cual le permite asignarlo como Sub- análisis en los APUS. |
| Devolutivo | Char | Configuración del insumo donde indica que es un insumo Devolutivo. |
| Stock Maximo | Int | Cantidad mínima en inventario. |
| Stock Minimo | Int | Cantidad máxima en inventario. |
| Valor Unitario | Money | Valor Unitario antes de IVA configurado en el maestro de insumos. |
| Porcentaje IVA | Float | Porcentaje de IVA configurado en el maestro de insumos. |
| Valor Neto | Money | Valor Unitario incluido IVA configurado en el maestro de insumos. |

**Tabla 3**  
*Tipos de Costo*

|  |  |  |
| --- | --- | --- |
| **Nombre Columna** | **Tipo De Dato** | **Descripción** |
| SkIdCapitulo | Int | Id de relación contra las otras tablas por medio del capitulo |
| SkIdEmpresa | smallint | Id de relación contra las otras tablas por medio de la empresa |
| SkIdProyecto | Int | Código del proyecto asociado al capitulo |
| Capitulo Numero | Int | Número del capitulo |
| Capitulo Descripcion | varchar | Descripción del capitulo |
| Tipo Costo | varchar | Descripción del tipo de costo |
| Tipo Costo Orden | smallint | Orden como se configuran los costos |
| Empresa | varchar | Nombre de la empresa que se encuentra asociado el capitulo |

**Tabla 4**  
*Control de Proyectos*

|  |  |  |
| --- | --- | --- |
| **Nombre Columna** | **Tipo De Dato** | **Descripción** |
| SkIdEmpresa | smallint | Id de relación para la tabla [ADP\_DTM\_DIM].[Empresa] para poder visualizar los datos demográficos de la empresa |
| Empresa | varchar | Nombre de la empresa que se encuentra asociado al control del proyecto |
| SkIdProyecto | int | Id de relación para la tabla [ADP\_DTM\_DIM].[Proyecto] para poder visualizar los datos demográficos del proyecto |
| SkIdFecha | int | Id de relación para la tabla [ADP\_DTM\_DIM].[Fecha] para poder visualizar los datos demográficos de la fecha de control del proyecto |
| SkIdClaseOrigen | smallint | Id de relación para la tabla [ADP\_DTM\_DIM].[ ControlCalseOrigen] para poder visualizar los datos demográficos de las clases y orígenes del control |
| SkIdInsumo | int | Id de relación para la tabla [ADP\_DTM\_DIM].[Insumo] para poder visualizar los datos demográficos de los insumos |
| SkIdCapitulo | int | Id de relación para la tabla [ADP\_DTM\_DIM].[CapituloPresupuesto] para poder visualizar los datos demográficos de los capítulos |
| SkIdItems | int | Id de relación para la tabla [ADP\_DTM\_DIM].[Items] para poder visualizar los datos demográficos de los items |
| Cantidad | numeric | Cantidad del control por proyecto |
| Valor Total | money | Valor total del control por proyecto |
| Origen Documento | bigint | Código origen del documento que tiene el control por proyecto |
| Origen Documento Detalle | int | Código origen del documento detallado que tiene el control por proyecto |
| Valor Sin IVA | money | Valor sin IVA que tiene el control por proyecto |

## Datasets

Dado que los datos originales están en diferentes tablas, se procede a realizar un proceso de ingeniería de datos para consolidar la información en un solo dataset, para ello se inicia con la lectura de los archivos en un notebook de la plataforma Databricks, luego se realiza joins entre las tablas, teniendo como base la tabla de hechos. Una vez realizado los joins, seleccionamos los campos que componen el dataset final. Los campos de la tabla son los siguientes:

**Tabla 4**  
*Dataset Final*

|  |  |  |
| --- | --- | --- |
| **Nombre Columna** | **Tipo De Dato** | **Descripción** |
| Str\_proyecto | varchar | Nombre del proyecto |
| Str\_vis | varchar | El proyecto es de tipo VIS (SI o NO) |
| Str\_insumo\_detalle | varchar | Descripción detallada asignada al insumo. |
| Str\_insumo\_agrupado | varchar | Descripción de la agrupación asignada al insumo. |
| Str\_tipo\_costo | varchar | Descripción del tipo de costo |
| Val\_total\_insumo | Int | Valor total del insumo |

Adicionalmente, los datasets se separarán en 80% para entrenamiento y 20% para validación.

## Analítica descriptiva

Se realizará un análisis descriptivo preliminar de los datos para comprender mejor su distribución y características principales, el cuál analizaremos de la siguiente manera:

* Validación de columnas del dataset
* Análisis de variables categóricas y numéricas
* Transformaciones de variables de entrada y salida
* Correlación entre variables
* Creación de datasets

**Validación de columnas**

Una vez terminado de organizar el dataset, iniciaremos con el proceso de validación de las columnas, para visualizar valores nulos, vacíos, y únicos y de esa manera iniciar a caracterizar y a limpiar la data. El resultado es el siguiente:

nulos por columna:

str\_proyecto: 0

str\_vis: 0

str\_insumo\_detalle: 0

str\_insumo\_agrupado: 0

str\_tipo\_costo: 4871

val\_total\_insumo: 0

-------------------------------------------

vacios por columna:

str\_proyecto: 0

str\_vis: 0

str\_insumo\_detalle: 0

str\_insumo\_agrupado: 0

str\_tipo\_costo: 0

val\_total\_insumo: 0

-------------------------------------------

valores únicos por columna:

str\_proyecto: 37

str\_vis: 2

str\_insumo\_detalle: 212

str\_insumo\_agrupado: 7

str\_tipo\_costo: 4

val\_total\_insumo: 7143

**Análisis de variables categóricas y numéricas**

Una vez realizada la evaluación por columna se procese a identificar cuáles son nuestras variables categóricas y numéricas el cuál va a permitir identificar la frecuencia de instancias entre variables.

#Listamos las variables categóricas

catCols = df\_dataset\_clean\_pd.select\_dtypes(include = ["object", 'category']).columns.tolist()

df\_dataset\_clean\_pd[catCols].head()

**Figura 1** *Lista de variables categóricas*

# A white rectangular object with black text Description automatically generated

# #Listamos de variables numéricas

numCols=df\_dataset\_clean\_pd.select\_dtypes(include = ['float64','int32','int64']).columns.tolist()

df\_dataset\_clean\_pd[numCols].head()

**Figura 2** *Lista de variables numéricas*

A screenshot of a table

Description automatically generated

Una vez definidas las variables procedemos a visualizar la frecuencia de las variables categóricas, y en este caso identificar cuales de los atributos que se encuentran en las variables influyen en el costo del proyecto de vivienda.

#Visualización de frecuencia de instancias para variables categóricas

multiple\_plot(2, df\_dataset\_clean\_pd , catCols, None, 'countplot', 'Frecuencia de instancias para variables categóricas',30)

**Figura 3** *Frecuencia de instancias de variables categóricas*

A comparison of different colored bars

Description automatically generated

De la Figura 3, podemos observar que el insumo que tiene más influencia en el proyecto son los materiales, asimismo, podemos identificar que el proyecto más costoso a sido el proyecto Marroco Torre 1.

Ahora se procede a visualizar las variables numéricas con respecto al costo del proyecto:

#Visualización de variables numéricas con respecto al Costo del Proyecto (Costo de Insumo)

multiple\_plot(1, df\_dataset\_clean\_pd , numCols, None, 'scatterplot', 'Relación entre las variables numéricas',30)

**Figura 4** *Relación entre variables numéricas*

A graph of a number of graphs

Description automatically generated with medium confidence

De la Figura 4 podemos identificar que existen valores muy distantes entre sí, lo cual implica realizar una identificación y eliminación de outliers para no dañar en un futuro al modelo.

A continuación, para la identificación de outliers graficamos una histograma y caja de bigotes.

#Histográma y caja de bigotes del Costo del Proyecto (Costo de Insumo)

plt.figure(figsize=(12,4))

plt.subplot(1,2,1)

plt.title('Real estate project Cost Distribution Plot')

sns.distplot(df\_dataset\_clean\_pd.val\_total\_insumo)

plt.subplot(1,2,2)

plt.title('Real estate project Cost Spread')

sns.boxplot(x=df\_dataset\_clean\_pd.val\_total\_insumo, palette='nipy\_spectral')

plt.show()

**Figura 5** *Histograma y caja de bigotes del Costo del Proyecto*

A comparison of a plot

Description automatically generated

Dado que ya se ha identificado los valores atípicos, ahora se procede a su eliminación:

#identificamos el outlier de acuerdo a la caja de bigotes

print(np.where(df\_dataset\_clean\_pd['val\_total\_insumo']>400000000))

# eliminamos el outlier

df\_dataset\_clean\_pd.drop([6147], axis=0, inplace=True)

df\_dataset\_clean\_pd.drop([13254], axis=0, inplace=True)

fig = px.box(df\_dataset\_clean\_pd, x="val\_total\_insumo",title='Boxplot for Real States Project Prices')

fig.show()

**Figura 6** *Caja de bigotes del Costo del Proyecto después de eliminar outliers*

A screen shot of a graph

Description automatically generated

**Transformaciones de variables de entrada y salida**

Ahora vamos a realizar transformaciones para que los valores de las variables sean insesgados, es decir, que los resultados sean parecidos a los reales. Para ello, elevaremos a la cubica la variable de salida, que pertenece al valor del proyecto con respecto al insumo:

#Tranformacion y = y^(1/3) Raiz cúbica

df\_dataset\_clean\_pd['val\_total\_insumo'] = np.cbrt(df\_dataset\_clean\_pd['val\_total\_insumo'])

#Histográma y caja de bigotes del Costo del Proyecto

plt.figure(figsize=(12,4))

plt.subplot(1,2,1)

plt.title('Real estate project Price Distribution Plot')

sns.distplot(df\_dataset\_clean\_pd.val\_total\_insumo)

plt.subplot(1,2,2)

plt.title('Real estate project Price Spread')

sns.boxplot(x=df\_dataset\_clean\_pd.val\_total\_insumo, palette='nipy\_spectral')

plt.show()

**Figura 7** *Histograma y caja de bigotes del Costo del Proyecto después de las transformaciones*

A comparison of a graph

Description automatically generated with medium confidence

Una vez finalizada la transformación de la variable de entrada podemos observar en la Figura 7, que los datos ya están más compactos, es decir, no se tiene valores atípicos y la densidad esta muy marcada en un sector.

Del mismo modo, se procede a realizar transformaciones a las variables de entrada:

# Tranformaciones X = X^2 a las variables de entrada

varTrans = ['str\_vis', 'str\_tipo\_costo']

for elemento in varTrans:

    df\_dataset\_clean\_pd[elemento]   = df\_dataset\_clean\_pd[elemento] \*\* 2

**Correlaciones**

Ahora procederemos a realizar una matriz de correlación para observar cómo se relacionan entre sí todos los posibles pares de valores de una tabla. La matriz de Correlación nos va a permitir resumir nuestro dataset y encontrar y mostrar patrones en ellos.

#Generación de la matriz de correlación

matrizCorr = df\_dataset\_clean\_pd.corr()

# Mapa de calor (Heatmap) de la matriz de correlaciones

fig, ax = plt.subplots(nrows=1, ncols=1, figsize=(10, 6))

sns.heatmap(

    matrizCorr,

    annot     = True,

    cbar      = False,

    annot\_kws = {"size": 8},

    vmin      = -1,

    vmax      = 1,

    center    = 0,

    cmap      = 'RdBu', #sns.diverging\_palette(20, 220, n=200),

    square    = True,

    ax        = ax

)

ax.set\_xticklabels(

    ax.get\_xticklabels(),

    rotation = 45,

    horizontalalignment = 'right',

)

ax.tick\_params(labelsize = 8)

**Figura 8** *Mapa de calor de la matriz de correlaciones*

A screenshot of a graph

Description automatically generated

#Matriz de correlación ordenada

pd.set\_option('display.float\_format', lambda x: '%.4f' % x)

#Creación de la matriz

df\_corr= tidy\_corr\_matrix(matrizCorr)

#Selección de las variables correlacionadas con el precio

df\_filtrado = df\_corr.loc[df\_corr['variable\_2'] == 'val\_total\_insumo']

#Se identifican las correlaciones más fuertes

df\_filtrado[df\_filtrado['abs\_r']>=0.001]

**Figura 9** *Valores de las correlaciones más fuertes*

A close-up of a web page

Description automatically generated

De acuerdo con la Figura 8 y 9, identificamos que la correlación es muy baja entre variables que estamos correlacionado, se debe buscar otras variables para una mejor correlación entre las variables y el precio, al menos debe ser mayor a 0.5.

**Creación de Datasets**

Adicionalmente, los datasets se separarán en 80% para entrenamiento y 20% para validación.

#Se crea una lista con todas las varibles de entrada

vDep = list(df\_dataset\_clean\_pd.columns)

vDep.remove('val\_total\_insumo')

#Se establece las variables de entrada 'X' y la variable de salida 'y'

X = df\_dataset\_clean\_pd.drop(columns ='val\_total\_insumo')

y = df\_dataset\_clean\_pd['val\_total\_insumo']

# División de los datos en train y test

#En este caso, el 80% de los datos se utilizarán para entrenar el modelo.

X\_train, X\_test, y\_train, y\_test = train\_test\_split(

                                        X,

                                        y.values.reshape(-1,1),

                                        train\_size   = 0.8,

                                        random\_state = 1234,

                                        shuffle      = True

                                    )

# 

# Referencias

American Psychological Association [APA]. (2020). *Publication Manual of the American Psychological Association* (7a ed.). American Psychological Association.

Context Constructora. 2024. ¿CUÁL ES LA DIFERENCIA ENTRE UNA VIVIENDA VIS Y NO VIS?. Medellín, CO.:Contex Publishing. Recuperado de <https://contex.com.co/vivienda-vis-y-no-vis/>

Scikit learn. 2024.linear\_model. USA.:Scikit Learng Developer Publishing. Recuperado de <https://scikit-learn.org/stable/supervised_learning.html>