| PROGRAMA DE FORMACIÓN | Gestión de analítica de datos |
| --- | --- |

| COMPETENCIA | 220501115 - Integrar datos según técnicas de visualización y metodologías de análisis. | RESULTADOS DE APRENDIZAJE | 220501115-02 Depurar la información acorde con los criterios técnicos y métodos estadísticos.  220501115-03 Procesar datos, de acuerdo a métodos estadísticos. |
| --- | --- | --- | --- |

| NÚMERO DEL COMPONENTE FORMATIVO | CF16 |
| --- | --- |
| NOMBRE DEL COMPONENTE FORMATIVO | Análisis y procesamiento de datos |
| BREVE DESCRIPCIÓN | Este componente maneja las bases del análisis y procesamiento de datos. Se presentan conceptos como limpieza de los datos, métodos estadísticos, manejo de los datos con Python en Google Colab y colocarlos dentro de un mapa para el filtrado de la información, utilizando gráficas dinámicas mediante la librería Plotly para una mejor comprensión y visualización con Power BI y Tableau. |
| PALABRAS CLAVE | Análisis, EDA, estadística, geoposicionamiento, procesamiento. |

| ÁREA OCUPACIONAL | Ciencias sociales, educación, servicios gubernamentales y religión |
| --- | --- |
| IDIOMA | Español |

**Tabla de contenido**

[**Introducción**](#_heading=h.3whwml4)

1. **Análisis** **de datos**

[**2 Modelos estadísticos**](#_heading=h.111kx3o)

[**3 Transformación de datos**](#_heading=h.3ygebqi)

1. **Introducción**

| Cuadro de texto |
| --- |
| Estimado aprendiz, reciba una cordial bienvenida a este recurso de aprendizaje, orientado al análisis y procesamiento de datos. Durante su desarrollo se abordan los ejes temáticos de: análisis de datos, modelos estadísticos y transformación de datos. Al finalizar el componente formativo, estará en capacidad de verificar la calidad de la información, comparar los datos, filtrarlos, ajustar contenidos de variables y elaborar un documento de depuración de información.  En el siguiente video conocerá, de forma general, la temática que se estudiará a lo largo del componente formativo. |

**Guion de video introductorio**

| **Tipo de recurso** | Video spot animado | | | |
| --- | --- | --- | --- | --- |
| **NOTA** | **La totalidad del texto locutado para el video no debe superar las 500 palabras aproximadamente** | | | |
| **Título** | Análisis y procesamiento de datos | | | |
| **Escena** | **Imagen** | **Sonido** | **Narración (voz en off)** | **Texto** |
| **1** | Tabla de bases de datos con almacenamiento de servidores y red en segundo plano del centro de datos  228131\_i1601 | NA | En el momento actual de conexión global la civilización produce una cantidad de datos enorme; los nuevos dispositivos permiten obtener información de las personas en tiempo real, sus hábitos de consumo, sus amistades, sus hobbies. | La civilización produce una cantidad de datos enormes, los nuevos dispositivos permiten obtener información de las personas en tiempo real. |
| **2** | estructura de organización y equipo simbolizada con cubos  228131\_i1602 | NA | Todo esto es de vital importancia para las empresas, pensado en la búsqueda de nuevos clientes, recomendación de productos, fidelización de clientes, mejoramiento respecto a la competencia y etc. | Es de vital importancia para las empresas pensado en búsqueda de nuevos clientes, recomendación de producto, fidelización de clientes. |
| **3** | Concepto de análisis de datos. Gente analizando las tablas de crecimiento. Análisis de datos empresariales. Aumentar las ventas y las habilidades. Análisis de personas. Monitorear la inversión. Gráfico de informes financieros. Gráficos y diagramas. Vector.  228131\_i1603 | NA | Las empresas buscan sacarle provecho a toda esa información en tiempo real con la finalidad de crecer, desarrollando políticas referentes a los datos como gobierno de datos, organizar, clasificar y guardar esa información en *data warehouse.* | Desarrollar políticas referentes a los datos como gobierno datos, organizar, clasificar, guardar esa información en data *warehouse*. |
| **4** | Businessman conectó dispositivos tecnológicos y aplicaciones de iconos con su teléfono móvil 3D representación  228131\_i1604 | NA | Por su parte, la información de diferentes fuentes como audio, video, redes sociales, se guarda en un *data lake* para futuro procesamiento, análisis e inferencia en la solución de problemas, | Audio, video, redes sociales, guardados en un *data lake* para futuro procesamiento, análisis, inferencia en la solución de problemas. |
| **5** | El concepto de IA (Inteligencia Artificial). Aprendizaje profundo. Aprendizaje automático. Singularidad.  228131\_i1605 | NA | teniendo presente el *machine learning* predictivo y la puesta de modelos en producción, utilizando la nube con alguno de los prestadores de servicio para ser consumidos por el usuario. | Utilizar la nube con alguno de los prestadores de servicio para ser consumidos por el usuario. |
| **6** | Panel de resultados de datos de KPI financieros en el equipo  228131\_i1606 | NA | La industria actual busca personal capacitado que comprenda todo el proceso que gira alrededor de los datos, su obtención, depuración, limpieza, procesamiento, modelo para ponerlo a producción; | La industria actual busca personal capacitado que comprenda todo el proceso que gira alrededor de los datos |
| **7** | Analista trabajando con Business Analytics y Data Management System en el equipo para realizar informes con KPI y métricas conectadas a la base de datos. Estrategia corporativa para finanzas, operaciones, ventas, marketingi  228131\_i1607 | NA | luego, el mantenimiento de ese modelo, su reentrenamiento, diferentes métricas o medidas del modelo y las pruebas o test sobre él. | luego el mantenimiento de ese modelo, reentrenamiento del modelo, diferentes métricas o medidas del modelo, las pruebas o test sobre el modelo. |
| **8** | 228131\_i1608 | NA | Hay una rama actual referente a verificación y mantenimiento de modelos que es *mlops*, basado en una integración continua e implementación de los modelos de *machine learning*. | *mlops* está basado en una integración continua e implementación de los modelos de *machine learning*. |
| **9** | Vista lateral de perfil de su buen barbudo con pantalones profesional experto en HTML la pantalla base de datos HTML en la estación de trabajo industrial de madera interior  228131\_i1609 | NA | El personal capacitado debe estar familiarizado con un lenguaje de programación, como *Python*, manejo de base de datos tanto *sql* como *no-sql (mongodb),* | El personal capacitado debe estar familiarizado con un lenguaje de programación, como *Python*, manejo de base de datos tanto *sq*l como *no-sql (mongodb),* |
| **10** | Mujer haciendo carga en la nube en la laptop en la oficina. Información de datos  228131\_i1610 | NA | procesos de *etl*, extraer, transformar y cargar la información en una base de datos, conocimiento de *machine learning* y cómo desplegar el modelo en algunos de los proveedores en la nube. | procesos de *etl,* extraer, transformar y cargar la información en una base de datos, conocimiento de *machine learning* y cómo desplegar el modelo en algunos de los proveedores en la nube. |
| **11** | Colaboración de equipo de negocios discutiendo el análisis de trabajo con datos financieros y el gráfico de informes de crecimiento de marketing en equipo, presentación e intercambio de ideas para la planificación de estrategias con beneficio de la empresa.  228131\_i1611 | NA | En este componente formativo se ubican y distinguen los conceptos y los términos más relevantes en la técnica de análisis exploratorio de datos, | Distinguir los conceptos y los términos más relevantes en la técnica de análisis exploratorio de datos, |
| **12** | Analista Jefe realiza presentación de reunión para un equipo de economistas. Muestra pizarra interactiva digital con análisis de crecimiento, gráficos, estadísticas y datos. La gente trabaja en una oficina creativa.  228131\_i1612 | NA | conocida como *Exploratory Data Analysis (EDA)*, para la fundamentación teórica en analítica de datos y para su aplicación de manera efectiva en la solución de problemas reales. | la fundamentación teórica en analítica de datos para su aplicación de manera efectiva en la solución de problemas reales. |
| **13** | Enseñanza electrónica Concepto de cursos en línea sobre tecnología de Internet.  228131\_i1613 | NA | La información se presenta de forma gradual iniciando por las bases en exploración de datos. La metodología es interactiva utilizando diferentes herramientas audiovisuales, con sus respectivas explicaciones teóricas; es importante el trabajo individual para maximizar la curva de aprendizaje. | Se desarrolla de forma gradual iniciando por las bases en exploración de datos. |
| **14** | Tablero, gran diseño para cualquier propósito del sitio. Plantilla de infografía empresarial. Ilustración plana vectorial. Concepto de gran tamaño UI de panel, diseño de plantilla de panel de administración de usuario de UX. Panel de administración de Analytics.  228131\_i1614 | NA | De igual manera, se trabaja la herramienta de nombre *Plotly* para generar visualizaciones con gráficos dinámicos o interactivos dentro de *Google Colab,* que permite ejecutar código en el lenguaje *Python* sin tener que instalar *Python.* | Se trabaja la herramienta de nombre *Plotly* para generar visualizaciones con gráficos dinámicos o interactivos dentro de *Google Colab.* |
| **15** | un proceso de garantía de calidad sin papel a mano utilizando un ordenador portátil con pantalla y documento virtuales y un concepto de gestión de ERP para su aprobación en línea.  228131\_i1615 | NA | Se utilizan los datos en formato *csv* tomados de *MEData* (Metada Alcaldía de Medellín) correspondiente a incidentes de tránsito ocurridos en esta ciudad. | Datos en formato *csv* tomados de *MEData* correspondiente a incidentes de tránsito ocurridos en Medellín. |
| **16** | POZNAN, POL - SEP 23, 2020: Laptop computer displaying logo of Power BI, a business analytics service by Microsofti  228131\_i1616 | NA | Se trabaja con *Power BI* y *Tableau* para diferentes visualizaciones de la información. | *Power BI*  y *Tableau* para diferentes visualizaciones de la información. |
| **17** | El tablero de datos empresariales proporciona análisis de inteligencia empresarial moderados para la planificación de estrategias de marketing  228131\_i1617 | NA | También se realiza un *dashboard* o tablero de mando usando *Power BI*, para una visualización interactiva de los datos, mediante filtrado de la información, con la finalidad de tener la fundamentación teórica. | Rrealizar un *dashboard* o tablero de mando usando *Power BI,* para una visualización interactiva de los datos. |
| **18** | Hombres de negocios mirando gráficos, hojas de cálculo, gráfico desarrollo financiero, cuentas bancarias, estadísticas, economía, análisis de datos, análisis de inversiones, bolsa de valores  228131\_i1618 | NA | Para la compresión del análisis estadístico, se utilizan diferentes gráficas histogramas, cajas y bigotes de forma dinámica. Asimismo, se utilizan términos como transformación de datos y métodos de discretización. | La compresión de análisis estadístico, utilizando diferentes gráficas histogramas, cajas y bigotes de forma dinámica. |
| **Nombre del archivo** | **228131\_v01** | | | |

**Desarrollo de contenidos**

1. **Análisis de datos**

| Cuadro de texto |
| --- |
| Recordemos que el análisis de datos es la ciencia que se encarga de examinar un conjunto de datos con el propósito de sacar conclusiones sobre la información y, de esta forma, poder tomar decisiones, o simplemente ampliar los conocimientos sobre diversos temas. |

**Técnicas de análisis exploratorio de datos (EDA)**

| **Tipo de recurso** | | Pestañas o tabs Verticales |
| --- | --- | --- |
| **Introducción** | | Es necesario tener claridad sobre los conceptos teóricos para llevarlos a la práctica. La finalidad de este componente es la de tener buenas bases estadísticas para el análisis de datos, mediante una técnica conocida como Análisis exploratorio de datos, EDA (Exploratory Data Analysis). |
| Imagen general que ilustre el tema    Imagen: 228131\_i1619 | | |
| ***EDA*** | **Análisis exploratorio de datos**  Consiste en tomar los datos en algunos de los formatos, por lo general CSV, separados por coma (,) o punto y coma (;) o cualquier otro símbolo de separación; también los datos podrían estar en formato de libro de excel .xlsx.  Se utiliza Python para la lectura de los datos, pasarlos a *data frame* y convertirlos a un formato tipo tabla, con filas y columnas.  Se trabaja con información real, datos de incidentes viales obtenidos de la página de datos abiertos de la Alcaldía de Medellín (Alcaldía de Medellín, 2022). | |
| ***Data frame*** | **Revisar la calidad de los datos del *data frame***  Información que se analiza dentro de GoogleColab. Después de pasar los datos a un *data frame,* se debe revisar la calidad de los mismos, como: datos repetidos, valores únicos, datos nulos, datos nan, datos atípicos y nombres de las columnas. Se observan los tipos de datos en cada columna, sin son *string*, objetos, numéricos, tipos de numéricos: enteros, flotantes, lógicos como booleanos cuyos valores son verdadero o falso y tipo fecha. Se revisa si es necesario convertirlos, como por ejemplo: de *string* a fecha o de *string* a numérico. | |
| **Convertir a formato fecha** | df['Fecha\_incidente']=pd.to\_datetime(df['Fecha\_incidente']), significa que mediante pandas con el alias o abreviación pd se utiliza el método *to\_datetime* para convertir la columna a formato fecha.  Otra forma de hacerlo es:  df["Fecha\_incidente"] = df["Fecha\_incidente"].astype("datetime64").  Al comando astype, internamente se le pasa el tipo de dato; en este caso, datetime64, podría ser flotante o entero si se necesita convertir un dato a flotante o a entero. | |
| **Valores únicos** | Para saber los valores únicos de una columna se utiliza: df['Sexo'].unique(), resultando en este caso ['M', 'F', 'Sin Inf'], M de masculino, F de femenino, Sin Inf de sin información.  Se puede tomar un subconjunto o filtrado de un *data frame* mediante:  df\_hombres = df[ df['Sexo'] == 'M' ], significa que se toma solo el género masculino y se coloca en una variable de nombre df\_hombres. | |
| **Valor *NULL*** | Indica algo que no existe y está vacío. Borrar valores NaN (*Not a Number*) significa No es un número. Ejemplo: df=df.dropna(), significa que borrará los datos NaN del data frame de nombre df. El valor *NULL* indica algo que no existe y está vacío diferente del *NAN*. | |
| **Dimensión del *data frame*** | Para saber la dimensión del *data frame,* como por ejemplo: df.shape, dará el número de filas y columnas del *data frame*. En los datos de incidentes de tránsito se obtiene (235845, 19), el primer elemento es el total de filas, el segundo elemento el total de columnas. | |
| **Analizar los datos** | **Analizar los datos categóricos: ordinales y nominales**  Categóricos: object  Numéricos: int64, significa valores enteros o discretos, representa un entero de 64 bits.  Numéricos: float64 (continuos), significa que toman valores reales de 64 bits.  En 64 bits, el rango es de -9.223.372.036.854.775.808 hasta 9.223.372.036.854.775.807.  Categórico ordinal: valores que representan categorías con alguna clasificación intrínseca, tiene características de orden como, por ejemplo: malo, bueno, excelente.  Categórico nominal: representa categorías sin clasificación intrínseca, ejemplo: las marcas de zapatos, género, raza, etnia.  Observar las primeras y las últimas filas del *data frame;* observar una columna en particular, también una fila en particular, tomar un número de filas del *data frame*. | |
| **Datos faltantes** | Hay técnicas para imputación de datos faltantes, por ejemplo, sustituir el dato por el promedio de la columna donde está ausente, por la mediana, todo va a depender del análisis detallado de los datos, cuántos datos faltantes hay en total; la última opción es eliminar la fila correspondiente a ese dato faltante. | |
| **Estadísticos descriptivos básicos** | Para una columna numérica, mediante df.describe() se obtiene:  El valor máximo, valor mínimo, promedio, desviación estándar, cuartil1 (Q1), cuartil2 (Q2), cuartil3 (Q3).  Contar los valores de una columna utilizando el método: value\_counts(), ejemplo: df['Comuna'].value\_counts(), significa que en la columna Comuna se obtendrá el número de datos por cada columna.  La Candelaria 36985, Castilla 22883, Laureles Estadio 19922, etc.  También se puede observar el porcentaje de esos valores referente al total, ejemplo:  (df['Clase\_incidente'].value\_counts() / len(df)).sort\_values(ascending=False)  Significa que el total de valores de la columna incidente se divide por la longitud del *data frame* y ordenado de mayor a menor. Para obtener el número de filas o longitud del *data frame,* se utiliza el comando len(). | |
| **Reemplazar símbolos** | Se puede reemplazar un símbolo en una columna mediante:  df\_limit['Latitud']=df\_limit['Latitud'].replace(',','.',regex=True), significa que en la columna Latitud, se reemplaza la coma por punto, para que no presente errores a nivel de cálculos.  Para cambiar la columna latitud a valor numérico mediante:  df\_limit['Latitud'] = pd.to\_numeric(df\_limit['Latitud'],errors = 'coerce')  Para obtener los tipos de datos del *data frame* se utiliza el método dtypes ejemplo:  df\_limit.dtypes, resultando Gravedad\_victima object, Año float64, Num\_dia float64, etc.  Para contar el total de valores nulos en el *data frame*, se le colocó el nombre df\_limit y se utiliza:  df\_limit.isnull().sum()  df.isnull() pregunta si hay valores nulos en el *data frame*. | |
| **Datos aleatorios** | Se pueden tomar datos aleatorios de un *data frame* en Python importando la librería *random* mediante *import random*, muestreo10 = df.sample(n=10), significa que toma una muestra aleatoria de 10 elementos del *data frame* llamado df y lo coloca en la variable de nombre muestreo10 que sería un *data frame* de 10 elementos.  Una selección aleatoria del 10 % de la tabla utilizando el comando simple, ejemplo:  pob\_10 = df.sample(frac =.1)  Se guarda el *data frame* con el 10 % del total en la variable pob\_10.  Si se requiere un 20 % se escribiría pob\_20 = df.sample(frac =.2).  Se guarda el *data frame* con el 20 % del total en la variable pob\_20. | |

| **Tipo de recurso** | Infografía interactiva 2 punto caliente | |
| --- | --- | --- |
| **Texto introductorio** | A continuación, se presentan los valores de la columna condición de forma porcentual. | |
|  | | |
| **Código de la imagen** | 228131\_i1620 | |
| **Punto modal 1** | **Muestra los comandos para escribir los valores de la columna condición en forma porcentual de mayor a menor.** | Punto 1 en amarillo |
| **Punto modal 2** | **Muestra los valores de la columna condición de forma porcentual de mayor a menor.** | Punto 2 en amarillo |

| Cuadro de texto |
| --- |
| Para familiarizarse con los mapas por medio de latitud y longitud, Observa los siguientes mapas de Colombia y Medellín, con los 100 puntos de los incidentes de tránsito ocurridos en esta ciudad.  También observa los gráficos dinámicos generados con la *libreria plotly.*  OJO COLOCAR LOS ARCHIVOS HMTL CORRESPONDIENTE AL MAPA DE COLOMBIA Y EL DE LOS ACCIDENTES EN MEDELLÍN.  Se encuentran en la carpeta CF16\_228131 - mapas  OJO COLOCAR LOS GRÁFICOS DINÁMICOS REALIZADOS CON LA LIBRERÍA PLOTLY cajas\_y\_bigotes\_Comunavrs.html, EdadvrsFrecuencia.html, cajas\_y\_bigotes\_Edad.html.  Se encuentran en la carpeta CF16\_228131 - Gráficos dinámicos PLOTLY |

**Herramientas para análisis y procesamiento de datos**

| Tipo de recurso | Cajón de texto de color |
| --- | --- |
| La profundización en el manejo de herramientas para el análisis y procesamiento de los datos, tiene como propósito mejorar la técnica conocida como Análisis exploratorio de datos, EDA (Exploratory Data Analysis), utilizando diferentes herramientas de procesamiento, análisis de datos y visualización. | |

| **Tipo de recurso** | Tarjetas avatar | |
| --- | --- | --- |
| **Introducción** | Entre las herramientas se encuentran Google Colab, Power BI, Tableau, CartoDb. | |
| **GoogleColab**  Se puede trabajar código de Python por celdas, importar librerías para visualización de datos como *matplolit, plotly, seaborn*, etc. | | Google Colaboratory Colab - Guía Completa Español - Marketing Branding  228131\_i1621 |
| **Power BI – Tableau Public**  Herramientas para visualización de datos Power BI y Tableau Public en las cuales se pueden hacer tableros de mando o *dashboard* y contar o narrar una historia sobre los datos o *storytelling*.  Este tipo de herramientas permite conectarse con diferentes tipos de datos y con datos en diferentes bases de datos, tanto *on premise* o en la nube, también se pueden publicar los *dashboard* y *storytelling* en la nube. | | POZNAN, POL - FEB 6, 2021: Laptop computer displaying logo of Power BI, a business analytics service by Microsoft  228131\_i1622 |
| **CartoDB**  Plataforma para almacenar y visualizar datos espaciales o geográficos; se debe tener latitud y longitud de los puntos para proceder a ubicarlos en el mapa. | | Mapping GDELT in CartoDB: A Tutorial – The GDELT Project  228131\_i1623 |

| Cuadro de texto |
| --- |
| A continuación, se explica un ejemplo paso a paso en la herramienta Power BI. Se trabajará con los datos de incidentes viales tomados de la página de datos abiertos de la Alcaldía de Medellín (Alcaldía de Medellín (2022). |

| **Tipo de recurso** | Slider Imagen | |
| --- | --- | --- |
| **Introducción** | Power BI permite obtener datos de diferentes fuentes de datos, analizarlos y presentar un análisis de estos a través de informes. A continuación, se explica secuencialmente su funcionamiento mediante ejemplos. | |
| En Power BI en el menú de inicio dar clic en el botón obtener datos, luego conjunto de datos de Power BI, después elige texto o CSV. | | .  228131\_i1624 |
| Se pueden transformar los tipos de datos por columna, convertir a texto, a número, fecha, etc. | | 228131\_i1625 |
| Las columnas latitud y longitud que están en formato texto o *string* se pueden cambiar a número. | | 228131\_i1626 |
| Muestra las columnas latitud y longitud en formato texto o *string,* se pueden cambiar a número donde está el símbolo número decimal. | | 228131\_i1627 |
| Muestra Las columnas latitud y longitud cambiadas a formato número. | | 228131\_i1628 |
| Los errores en las columnas se pueden eliminar dando clic derecho en “Quitar errores”. | | 228131\_i1629 |
| Se puede observar la columna latitud sin errores. | | 228131\_i1630 |
| Se pueden reemplazar valores en una columna. | | 228131\_i1631 |
| Las columnas se pueden filtrar de forma ascendente, descendente, por valores, etc. | | 228131\_i1632 |
| En Power BI se puede observar el modelo. | | 228131\_i1633 |
| En Power BI se pueden crear un *dashboard* o tablero de mando luego de hacer la limpieza de los datos que consiste en borrar los datos duplicados, los valores nan, los valores nulos, el tipo de variable correcto para cada columna, etc. | | Imagen: 228131\_i1634 |
| En la configuración interna del computador se puede cambiar a región español (Latinoamérica) para que Power BI reconozca el decimal como un punto y la coma para separación de mil**.** | | Imagen: 228131\_i1635 |
| Muestra la siguiente pestaña para cambiar a región español (Latinoamérica). | | Imagen: 228131\_i1636 |
| Muestra la última pestaña para cambiar a región español (Latinoamérica) | | Imagen: 228131\_i1637 |
| Si el mapa está deshabilitado debe ir a archivo-> opciones y configuración-> opciones ->global ->seguridad, por último, marcar uso de elementos visuales de mapa y mapa coroplético. | | Imagen: 228131\_i1638 |
| Muestra cómo habilitar el mapa. | | Imagen: 228131\_i1639 |
| Para el *dashboard* se escoge de la visualización la figura tabla y, luego, las columnas, en este caso año, mes, día, clase\_incidente, barrio, comuna, grupo\_edad, condición. | | Imagen: 228131\_i1640 |
| En el *dashboard* se escoge el gráfico de anillos donde está el recuento por comuna referente a la clase de incidente (choque, otro, atropello, caída del ocupante, volcamiento). | | Imagen: 228131\_i1641 |
| Filtrado por las columnas, en este caso: año, mes, día, clase\_incidente, barrio, comuna, grupo\_edad, condición. | | .  Imagen: 228131\_i1642 |
| En el *dashboard* final se muestran dos tarjetas donde está la hora del incidente y la condición. | | Imagen: 228131\_i1643 |

| Cuadro de texto |
| --- |
| En el ejemplo anterior, se presentó el funcionamiento de Power BI y la creación de tableros dinámicos en un proceso de análisis de datos reales sobre incidentes de tránsito en la ciudad de Medellín. A continuación, se ilustra un breve ejemplo del uso de la herramienta Tableau Public. |

| **Tipo de recurso** | Slider presentación | |
| --- | --- | --- |
| **Introducción** |  | |
| **Tableau Public**  Es otra herramienta para la visualización y el análisis de datos. Al conectar, se escoge el tipo de fuente de los datos, en este caso: archivo de texto. | | Imagen: 228131\_i1644 |
| En el archivo de texto se seleccionan los datos de incidentes de tránsito en Medellín. | | Imagen: 228131\_i1645 |
| Se puede observar una vista previa de los datos en Tableau. | | Imagen: 228131\_i1646 |
| Al hacer clic en la hoja 1, en esa región se pueden hacer las diferentes visualizaciones o gráficos. | | Imagen: 228131\_i1647 |
| En las columnas, se encuentran escritos diferentes símbolos. ABC significa que la columna es de tipo texto o *string*, símbolo de calendario referente a una fecha, símbolo de numeral referente a un valor numérico.  Para latitud y longitud, se cambia de número a función geográfica latitud y longitud, función geográfica longitud. Queda con un símbolo de globo terráqueo o mundo. | | Imagen: 228131\_i1648 |
| Ahora, hacer clic en actualizar, para ver los datos en las diferentes columnas. | | .  Imagen: 228131\_i1649 |
| En la columna se arrastra la variable año, en la fila se arrastra la variable grupo\_edad y en texto se arrastra la variable edad, quedando etiquetada por edad. | | Imagen: 228131\_i1650 |

**Criterios de respaldo y recuperación de información**

| **Tipo de recurso** | Tarjetas animadas |
| --- | --- |
| **Introducción** | La información que genera una empresa u organización debe almacenarse en una base de datos ***on premise*** o en la nube y tener como respaldo una copia de seguridad llamada *backups*. Las copias de seguridad deben realizarse periódicamente, dependiendo de la rapidez en la generación de datos y de la actividad de la empresa. |
|  | Los *backups* se pueden programar mediante un *trigger* o disparador, a un cierto horario y, generalmente, se realiza en la madrugada. |
| Imagen: 228131\_i1651 | **Servicio en la nube**  Amazon Web Services (Aws), Google Cloud Platform (GCP), Microsoft Azure, IBM Cloud, etc.  Dependiendo de la frecuencia de acceso a la información de respaldo o *backups*, se pueden utilizar varios planes para disminuir costos. |
| Imagen: 228131\_i1652 | **Tipos de copias de seguridad**  Hay diferentes tipos de copias de seguridad: completa, incremental, diferencial, espejo, sintética completa, etc. Por ejemplo, la completa incluye todos los archivos y carpetas seleccionados. |
| Imagen: 228131\_i1653 | **Propósito del *backup***  Respaldar la información generada para prevenir la pérdida total o parcial de la misma, como consecuencia de algún evento. |

**Gestión de seguridad de la información**

| **Tipo de recurso** | Rutas / Pasos. Verticales 1 |
| --- | --- |
| **Introducción** | Existen diferentes proveedores de servicios que prestan esa asistencia en la nube como: Amazon Web Services (Aws), Google Cloud Platform (GCP), Microsoft Azure, IBM Cloud, etc. |
| Imagen: 228131\_i1654 | |
| **Botón 1** | Se pueden hacer réplicas de la información para que no esté centralizada en una sola región del servidor. Esto genera un mayor costo, pero también tranquilidad en caso de alguna contingencia de orden mayor. |
| **Botón 2** | Es recomendable que la región escogida esté cerca de la información para mejorar la latencia o tiempo de respuesta a una petición realizada por un usuario. |
| **Botón 3** | Para acceder a las instancias en la nube se debe tener una clave cifrada, generada por el proveedor de servicio. |
| **Botón 4** | La seguridad debe ser tanto interna como externa. En la parte interna, se deben crear los grupos de seguridad muy bien definidos, separarlos dependiendo del área en la que trabajan en la empresa y asignarles los permisos necesarios para acceder a la información, según las funciones que desempeñan. Ejemplo: dependiendo del tipo de empresa, se crean diferentes grupos como: administrativos, contabilidad, financiera, ventas, mercadeo, desarrollo, producción, etc. |
| **Botón 5** | Se deben dar los menores permisos a los usuarios, es decir, solo los necesarios dependiendo de su grupo de trabajo.  Si el empleado trabaja desde casa, se debe asesorar para que lleve un estricto protocolo de acceso a la información de la empresa desde su portátil. |
| **Botón 6** | Las contraseñas se deben guardar de forma segura para evitar el acceso a la información de personas fuera de la empresa. Es importante tener diferentes claves cifradas aleatorias para cada servicio. Se debe tener un gestor de contraseñas con autenticación en dos factores y con un límite de tiempo de vencimiento de la clave, por ejemplo, de 30 segundos. |

1. **Modelos estadísticos**

| Cuadro de texto |
| --- |
| Cuando se habla de métodos o modelos estadísticos para el análisis de datos, se indica que estos consisten en una secuencia de procedimientos, que permiten el manejo de los datos de investigación, sean estos cualitativos o cuantitativos. |

**Tipos de variables**

| Cuadro de texto |
| --- |
| Hemos avanzado en el manejo de herramientas para la creación de visualizaciones. Ahora, se continúa con el desarrollo de los temas relacionados con tipos de variables cualitativas o categóricas y cuantitativas o numéricas, entre ellas las discretas y las continuas. |

| **Tipo de recurso** | Slider imagen | |
| --- | --- | --- |
| **Introducción** | A continuación, se explican los diferentes tipos de variables, mediante teoría-práctica, manejando un ejemplo real utilizando gráficos, generados dentro de Google Colab, utilizando el lenguaje Python.  Para la ejemplificación, se tomaron pantallazos de los datos de incidentes de tránsito obtenidos de la página de datos abiertos de la Alcaldía de Medellín en formato CSV (Alcaldía de Medellín, 2022). | |
| Categóricos: *objec*t y *booleano* (*booleanos* o lógicos, verdadero o falso, *true, false*).  Numéricos: int64, significa valores enteros o discretos, representa un entero de 64 bits.  Numéricos: float64 (continuos), significa que toman valores reales de 64 bits.  En 64 bits, el rango es de -9.223.372.036.854.775.808 hasta 9.223.372.036.854.775.807  df.info() da información sobre el tipo de variables en las columnas del *data frame*. También df.dtypes, nos da información sobre las variables. | | Imagen: 228131\_i1655 |
| La columna “Gravedad\_victima” es una variable de tipo objeto no nula.  La columna “Edad” es una variable de tipo objeto no nula: 17, 20, 18,19, 39.  df.nunique() permite saber los valores únicos que toma la columna.  La columna “Gravedad\_victima” tiene 2 valores únicos los cuales son “Heridos y Muertos”. | | Imagen: 228131\_i1656 |
| La columna sexo o género es una variable categórica nominal, toma los valores: masculino y femenino.  La columna clase de incidente es una variable categórica nominal, toma los valores: choque, otro, atropello, caída ocupante, volcamiento, incendio. | | Poner la imagen anterior  Imagen: 228131\_i16xx |
| La columna “num\_dia” es una variable numérica discreta, toma los valores: 1.0, 2.0, 3.0, 4.0  Las variables cuantitativas son continuas y discretas.  Ejemplo de variables cuantitativas continuas: el peso, la edad tomando los meses, días.  Las variables cualitativas son categóricas ordinales y categóricas nominales.  Las variables categóricas nominales son las que representan categorías sin clasificación intrínseca, ejemplo: las marcas de zapatos, género, raza, etnia. | | Poner la imagen anterior  Imagen: 228131\_i16xx |
| Las variables categóricas ordinales son valores que representan categorías con alguna clasificación intrínseca, tienen características de orden, ejemplo: malo, bueno, excelente. | | Poner la imagen anterior  Imagen: 228131\_i1657 |

* 1. **Muestreo estadístico**

| **Tipo de recurso** | | Pestañas o tabs Verticales |
| --- | --- | --- |
| **Introducción** | | En estadística una muestra es un subconjunto de casos o individuos en una población; en varias aplicaciones, es importante que esta sea representativa y, por esta razón, se debe elegir un método de muestreo apropiado para crear una muestra suficientemente aleatoria. |
| Imagen: 228131\_i1658 | | |
| **Muestreo aleatorio simple** | Técnica de muestreo donde todos los individuos que conforman la muestra tienen la misma probabilidad de ser elegidos. Si el individuo puede ser seleccionado más de una vez, se está en presencia de un muestreo aleatorio simple con reposición; en caso contrario, sería un muestreo aleatorio simple sin reposición.  **Ejemplo**  Se tiene una población de 500 personas y se quiere sacar una muestra de 50. Se asigna a las 500 personas un número y, de la misma forma que un sorteo de lotería, se seleccionan 50 números al azar. La totalidad de estos números es la muestra de 50 personas. | |
| **Muestreo aleatorio estratificado** | Se utiliza cuando se quiere saber la relación entre los subgrupos o estratos por alguna característica o múltiples características en común; como el año cuando ocurrió el incidente de tránsito, la edad de la persona involucrada, etc.  **Ejemplo**  Dividir la muestra de las personas adultas menores o iguales a 52 años que tuvieron un incidente de tránsito en Medellín en 7 subgrupos por edad, cada letra corresponde a un grupo: a) 18-25 años, b) 26-33 años, c) 34-41 años, d) 42-49 años, e) 43-50 años, f) 44-51 años, g) 45-52 años. | |
| **Ejemplo muestreo estratificado** | La población de 204185 personas se puede dividir por estratos según el año del incidente de tránsito. La característica o estrato sería el año del incidente. Tomando un subgrupo o subpoblación por año, se debe realizar el muestreo aleatorio en cada estrato. Por ejemplo, tomar los que tuvieron incidentes en el 2014 que son 31411 y de allí, tomar la muestra aleatoria con un tamaño representativo determinado. En ese año ocurrió el mayor número de incidentes, un total de 31411.   | Años | Total de incidentes de tránsito | | --- | --- | | 2014 | 31411 | | 2015 | 32622 | | 2016 | 33791 | | 2017 | 31658 | | 2018 | 29082 | | 2019 | 31876 | | 2020 | 23676 | | 2021 | 21727 | | |
| **Muestreo sistemático** | Se utiliza cuando el muestreo aleatorio simple no se puede llevar a cabo por algún motivo.  **Ejemplo:** de una población de 235845 se quiere una muestra de 500 elementos o individuos, para ello se divide el total de la población en 50 pedazos o fragmentos aleatorios de 10 individuos. Luego se selecciona un número al azar entre 1 y 10.  Elegir de forma aleatoria el primer individuo, luego definir el tamaño de un intervalo para seleccionar los otros individuos que hacen parte de la muestra. Este tamaño del intervalo lo define la persona a cargo de la investigación. | |
| **Ejemplo muestreo sistémico** | Se va a extraer una muestra de 10 personas de una muestra de 50 personas. El primer individuo seleccionado para la muestra es el número 5 con un tamaño o intervalo de 4, decidido por el investigador.  Los elementos de la muestra serían: 5, 9, 13, 17, 21, 25, 29, 33, 37, 41.  En Google Colab se puede hacer el muestreo aleatorio simple, importando la librería *random,* relacionada con valores aleatorios, se carga mediante *import random.*  muestreo15= df.sample(n=15), significa que de la *data frame* llamado df toma una muestra aleatoria de 15 elementos de la población 235845, se guarda en la variable muestreo15.  Muestreo aleatorio simple    Imagen: 228131\_i1659 | |

| Tipo de recurso | Cajón de texto de color |
| --- | --- |
| De la población 235845 se puede tomar un porcentaje aleatorio, por ejemplo, un 10 %.  pob\_10 = df.sample(frac =.1), el valor .1 significa 10 % se guarda en la variable pob\_10.  Muestreo aleatorio    Imagen: 228131\_i1660 | |

* 1. **Análisis bivariado**

| Tipo de recurso | Cajón de texto de color |
| --- | --- |
| En el análisis bivariado, se busca observar la relación entre dos variables cuantitativas continuas.Un joven sostiene una lupa en la mano, mirando cifras estadísticas - un concepto de negocio sobre el tema de muestreo y selección del grupo objetivo adecuado y examinándolo  Con la librería seaborn de Python, con el alias o abreviación sns, se puede hacer análisis bivariado, es decir, de 2 variables, con la finalidad de observar la relación entre las dos o el comportamiento de una variable en relación con la otra. Teniendo como referencia los incidentes ocurridos en Medellín, se puede analizar, por ejemplo, la clase de incidente, en comparación con el año, y relacionarlo con el número de días.  Se debe observar cómo los datos están distribuidos, si tienen forma de distribución normal o *Campana de Gauss*.  Imagen: 228131\_i1661 | |

| **Tipo de recurso** | Rutas / Pasos. Verticales 1 |
| --- | --- |
| **Introducción** | En el análisis bivariado se puede utilizar el diagrama de dispersión, coeficiente de correlación y la regresión lineal simple. |
| Imagen: 228131\_i1662 | |
| **Botón 1** | Diagrama de dispersión el objetivo es averiguar la relación entre las dos variables numéricas. |
| **Botón 2** | El coeficiente de correlación de *Pearson* mide la relación entre dos variables continuas, tiene un rango de valores entre -1 y 1. |
| **Botón 3** | Correlación positiva si una variable aumenta la otra también. Si una variable disminuye la otra también. |
| **Botón 4** | Correlación negativa si una variable aumenta la otra disminuye. Si una variable disminuye la otra aumenta. |
| **Botón 5** | Si el coeficiente de correlación de Pearson es 1 se habla de correlación perfecta.  Si el coeficiente de correlación de Pearson es 0 se habla de correlación nula.  Si el coeficiente de correlación de Pearson está entre 0.8 y 1, sin incluir los extremos, se habla de correlación muy alta.  Si el coeficiente de correlación de Pearson está entre 0.6 y 0.8, sin incluir los extremos, se habla de correlación alta.  Si el coeficiente de correlación de Pearson está entre 0.4 y 0.6, sin incluir los extremos, se habla de correlación moderada.  Si el coeficiente de correlación de Pearson está entre 0.4 y 0.2, sin incluir los extremos, se habla de correlación baja.  Si el coeficiente de correlación de Pearson está entre 0 y 0.2, sin incluir los extremos se habla de correlación muy baja. |
| **Botón 6** | Utilizando *joint* plot se pueden hacer simultáneamente la gráfica de las dos variables. En la parte superior se observa el gráfico sobre cómo está distribuida una variable y en la parte derecha, el gráfico de la distribución de la otra variable.    Imagen: 228131\_i1663 |

* 1. Regresiones

| Tipo de recurso | Cajón de texto de color |
| --- | --- |
| La presente sección se trabaja con la conocida *data* de precios de casas en Boston, Estados Unidos; para ello ver el documento Jupyter notebook en el material para descargar de nombre regresion.ipynb, donde se maneja tanto la regresión lineal simple, como la regresión lineal múltiple.Fondo del gráfico comercial: Análisis de la contabilidad de negocio en hojas de información. Empresario trabajando en el análisis de cartas de inversión para el mercado del oro, el mercado de divisas y el mercado del petróleo crudo.  En la regresión lineal simple se tiene dos variables: “y” la que se va a predecir, llamada variable dependiente y tenemos también la “x”, llamada variable independiente.  Se podría escribir como una ecuación lineal de la forma y = m\*x +b  Donde m es la pendiente o inclinación de la recta, b es el punto de corte con el eje y.    En la regresión lineal múltiple se tiene una sola variable a predecir “y”, también varias variables independientes:  Las constantes:  La ecuación se puede escribir de la forma:  Y=  Imagen: 228131\_i1664 | |

| Cuadro de texto |
| --- |
| Para reforzar los conceptos sobre regresiones, se sugiere consultar el libro *Estadística descriptiva regresión y probabilidades con aplicaciones****,*** el cual se encuentra en el material complementario. |

* 1. **Corroboración de hipótesis**

| **Tipo de recurso** | Acordeón tipo 2 |
| --- | --- |
| **Introducción** | Consiste en contrastar dos hipótesis opuestas, sobre algún caso de estudio, llamadas hipótesis nula (Ho) e hipótesis alternativa (H1).  Por ejemplo, se quiere corroborar si hay alguna diferencia significativa entre el promedio de incidentes de tránsito en las mujeres respecto a la de los hombres. En este caso sería una diferencia de medias. |
| Imagen: 228131\_i1665 | |
| **Promedio de mujeres que tienen incidentes de tránsito**  El promedio de las mujeres que tienen incidentes de tránsito sería: y el promedio de los hombres que tienen incidentes de tránsito sería:  Para ello se propone una prueba de hipótesis:  Las dos medias son iguales , pasando a la izquierda resulta  Las dos medias son diferentes , pasando a la izquierda resulta  **Resumen:**  Ho = hipótesis nula =  H1 = hipótesis alternativa = | |
| **Error tipo 1**  En estadística cuando se plantea una prueba de hipótesis debemos hablar de un error que se puede cometer, se llama **error tipo 1 ( ocurre cuando se rechaza la hipótesis nula siendo esta verdadera.** El error tipo 1( se conoce como falso positivo. | |
| **Intervalo de confianza**  Con un intervalo de confianza del 95 %, el resto sería el error, es decir, el error tipo 1 es = 5 %= 0.05, como se está diciendo que son diferentes, no mayor ni menor, se debe trabajar con 2 colas el error se divide entre las dos colas  **.** El valor **z** en la **distribución normal** **estandarizada** correspondiente a ese Intervalo de Confianzadel **95 % es Z=1.96** | |
| **Calcular el estadístico**  Calculará el **estadístico**, si ese estadístico cae en el intervalo de z **(-1.96, 1.96)** **significa que cae en la región de aceptación de Ho,** por lo tanto, podemos decir que con un 95 % de confianza no hay diferencia significativa entre los hombres y las mujeres que tienen incidentes de tránsito. | |
| **Aceptación o rechazo de hipótesis**  Pero si el **estadístico** cae fuera del intervalo de z (-1.96, 1.96), es decir, valores mayores a z = 1.96 o valores menores a z = -1.96 **significa que cae en la región de rechazo de Ho,** por lo tanto, aceptación de la hipótesis alternativa H1, podemos afirmar que con un 95 % de confianza, que si hay diferencia significativa entre los hombres y las mujeres que tienen incidentes de tránsito. | |

1. **Transformación de datos**

| Cuadro de texto |
| --- |
| La transformación de datos es el proceso de convertir datos sin procesar de un formato a otro, para que el sistema o la aplicación de destino los pueda utilizar. Incluye múltiples actividades, como 'transformar' sus datos filtrándolos según ciertas reglas y uniendo diferentes campos para obtener una vista consolidada. |

* 1. **Proceso de transformación de datos**

| Tipo de recurso | Cajón de texto de color |
| --- | --- |
| Esta técnica consiste en tomar los datos en algunos de los formatos generalmente CSV, separados por comas (,) o punto y comas (;) o cualquier otro símbolo de separación; también podrían estar en formato de libro de excel .xlsx. modificarlos para un correcto análisis de los datos. New york, USA - August 18, 2017: Microsoft excel menu on laptop screen close-up  Se utilizará Python para la lectura de los datos, pasarlos a *data frame* y convertirlo a un formato tipo tabla con filas y columnas.  Se trabaja con información real, datos sobre incidentes de tránsito en Medellín obtenidos de la página de datos abiertos de la Alcaldía.  Imagen: 228131\_i1666 | |

| **Tipo de recurso** | Slider Presentación | |
| --- | --- | --- |
| **Introducción** | La finalidad de este componente es tener buenas bases estadísticas para el análisis de datos, utilizando la técnica Análisis exploratorio de datos, EDA (Exploratory Data Analysis). | |
| En el proceso de transformación de datos se quiere borrar los datos NaN de una columna, por ejemplo, de la columna Edad. | | Imagen: 228131\_i1667 |
| Se quiere saber si hay algún valor nulo en la columna o variable Edad. | | Imagen: 228131\_i1668 |
| La fecha está en formato objeto o *string* y se puede cambiar a formato fecha, como se observa en la figura: Muestra los tipos de columnas.  Utilizando:  df['Fecha\_incidente']=pd.to\_datetime(df['Fecha\_incidente'])  Se puede hacer de otra forma utilizando:  df["Fecha\_incidente"] = df["Fecha\_incidente"].astype("datetime64") | | Imagen: 228131\_i1669 |
| Se pueden saber los tipos de variables de las columnas utilizando el método df.dtypes. | | Imagen: 228131\_i1670 |
| Cambiar la latitud y longitud de objeto a numérico. | | Imagen: 228131\_i1671 |
| Cambiar la coma por el punto en las variables latitud y longitud, | | Imagen: 228131\_i1672 |

* 1. **Métodos de discretización**

| Tipo de recurso | Cajón de texto de color |
| --- | --- |
| La finalidad del método de discretización es la de convertir una variable numérica en variable categórica.comportamiento de las multitudes que miden las estadísticas de muestreo social experimentan la investigación de la sociedad  Se requiere una variable numérica continua, para convertirla en categórica; este proceso se llama **discretización de variables**. Por ejemplo, la edad que es una continua numérica, se puede dividir por categorías o intervalos, de acuerdo con la edad. Las categorías podrían ser niños, jóvenes, adultos, ancianos. Una de las aplicaciones de este método es la de convertir un proceso de regresión, a clasificación. Se trata de predecir el valor exacto de la edad y determinar si esta variable corresponde a un niño, joven, adulto, o anciano. Por el cambio del problema de regresión a clasificación, se cambiaría a las técnicas de *Machine Learning* que se pueden trabajar, debido a que se modificó el enfoque del problema. También se podría utilizar para la clasificación de valores atípicos, valores que son muy altos o bajos quedarían dentro de su categoría.  Imagen: 228131\_i1673 | |

| **Tipo de recurso** | Slider Imagen | |
| --- | --- | --- |
| **Introducción** | Se utiliza la librería de Pandas *de* Python para discretizar variables. | |
| Para la variable Edad este tipo objeto o *string* se debe convertir primero a numérica antes de discretizar, utilizando la librería de pandas con el alias o abreviación pd.  df['Edad'] = pd.to\_numeric(df['Edad'],errors = 'coerce') | | Imagen: 228131\_i1674 |
| Categorizar la variable edad:  Niño = [0,10]  Joven = (10,18]  Adulto = (18,60]  Anciano = (60, inf]  Inf=infinito  Tenemos 4 categorías o intervalos de diferentes tamaños y se utilizarán en los datos de incidentes viales en la ciudad de Medellín.  Pandas con su alias pd, pd.cut( ) para dividir o cortar en los intervalos, trabajando con las etiquetas labels=["niño","joven","adulto","anciano"] que corresponden a cada intervalo. np.inf es el número infinito. | | Imagen: 228131\_i1675 |

| **Tipo de recurso** | Infografía interactiva 1 Punto caliente | |
| --- | --- | --- |
| **Texto introductorio** | A continuación, se presenta la discretización de una variable por categorías. | |
| **Imagen**      Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico, Teams  Descripción generada automáticamente | | |
| **Código de la imagen** | Imagen: 228131\_i1676 | |
| **Punto caliente 1** | Muestra la discretización de una variable por categorías. | P1 en amarillos |
| **Punto caliente 2** | Muestra la salida de la discretización. | P2 en amarillo |

| Cuadro de texto |
| --- |
| Para poner en práctica los conocimientos abordados en este proceso de aprendizaje, se sugiere descargar los archivos que se usaron para realizar la práctica.  SE ENCUENTRAN EN LA CARPETA CF16\_228131-Material para Descargar |

| Cuadro de texto |
| --- |
| Recuerde explorar los demás recursos que se encuentran disponibles en este componente formativo; para ello, diríjase al menú principal, donde encontrará la síntesis, una actividad didáctica para reforzar los conceptos estudiados, material complementario, entre otros. |

**Síntesis**

| **Tipo de recurso** | Síntesis |
| --- | --- |
|  | |
| **Introducción** | El siguiente mapa integra los criterios y especificidades de los conocimientos expuestos en el presente componente formativo. |
| Imagen: 228131\_i1677 | |

**Actividad interactiva**

| **Tipo de recurso** | Actividad didáctica. Verdadero o falso | |
| --- | --- | --- |
| El propósito esta actividad es verificar el grado de conocimientos obtenido en el desarrollo de los contenidos de aprendizaje del componente formativo. Lea atentamente cada afirmación y señale si es verdadera o falsa. | | Imagen: 228131\_i1678 |
| * + - 1. El comando df.shape, sirve para reemplazar un valor del *data frame*. | | Imagen: 228131\_i1679 |
| **Verdadero** | | **Falso** (correcto) |
| **Retroalimentación**: para saber la dimensión del *data frame* ejemplo: df.shape, dará el número de filas y columnas del data *frame*. | | |
| * + - 1. El comando df.dropna(), sirve para saber los valores Nan del *data frame*. | | Imagen: 228131\_i1680 |
| **Verdadero** | | **Falso** (correcto) |
| **Retroalimentación:** el comando df.dropna(), significa que borrará los datos NaN del *data frame* de nombre df. | | |
| * + - 1. El comando df.dtypes, sirve para cambiar el tipo de dato. | | Imagen: 228131\_i1681 |
| **Verdadero** | | **Falso** (correcto) |
| **Retroalimentación:** el comando df.dtypes sirve para saber los tipos de datos que tienen las columnas del *data frame* de nombre df, estos datos pueden ser *string*, objetos, numéricos, tipos de numéricos: enteros, flotantes. Lógicos como booleanos cuyos valores son verdadero o falso en ingles *true* o *false*. Tipo fecha date. | | |
| * + - 1. El comando df.unique(), sirve para cambiar los datos a valores únicos. | | Imagen: 228131\_i1682 |
| **Verdadero** | | **Falso** (correcto) |
| **Retroalimentación:**df.unique(), sirve para saber los valores únicos de una columna, ejemplo si la columna es género daría masculino y femenino. | | |
| * + - 1. El comando df.value\_counts (), sirve para saber el número de filas del *data frames.* | | Imagen: 228131\_i1683 |
| **Verdadero** | | **Falso** (correcto) |
| **Retroalimentación**: el comando df.value\_counts ()*,* cuenta los diferentes valores de una columna, ejemplo si la columna es género daría femenino 50, masculino 40. | | |
| * + - 1. El comando len(df)), sirve para saber la longitud del *data frame*. | | Imagen: 228131\_i1684 |
| **Verdadero (correcto)** | | **Falso** |
| **Retroalimentación:** el comando len(df)), es usado para saber la longitud del *data frame*, también para saber la longitud de una variable. | | |
| * + - 1. El comando *df.isnull(),* sirve para saber si hay valores nulos en el *data frame.* | | Imagen: 228131\_i1685 |
| **Verdadero (correcto)** | | **Falso** |
| **Retroalimentación:** el comando df.isnull(),es utilizado para saber si hay valores nulos en el *data frame*. Las respuestas son *true* (si), *false* (no). | | |
| * + - 1. Importando la librería random, el comando df.sample(frac =.1), sirve para obtener una muestra aleatoria de un 10 % del *data frame*. | | Imagen: 228131\_i1686 |
| **Verdadero (correcto)** | | **Falso** |
| **Retroalimentación:** el comando df.sample(frac =.1), es utilizado para obtener una muestra aleatoria de un 10 % del *data frame*, si se quiere una muestra aleatoria del 20 % sería df.sample(frac =.2). | | |
| * + - 1. Importando la librería random, el comando df.sample(n=10), sirve para tomar una muestra aleatoria de 10 elementos del *data frame*. | | Imagen: 228131\_i1687 |
| **Verdadero (correcto)** | | **Falso** |
| **Retroalimentación**: el comando df.sample(n=10), es utilizado para tomar una muestra aleatoria de 10 elementos del *data frame*, si se quiere una muestra aleatoria de 15 elementos sería df.sample(n=15). | | |
| * + - 1. El comando df.info() da información sobre el tipo de variables en las columnas del *data frame*. | | Imagen: 228131\_i1688 |
| **Verdadero (correcto)** | | **Falso** |
| **Retroalimentación:** df.info() da información sobre el tipo de variables en las columnas del *data frame*, ejemplo variables tipo: objeto, *string*, enteras, flotante, etc. | | |

**Material complementario**

| Tipo de recurso | Material complementario | | |
| --- | --- | --- | --- |
| Tema | Referencia APA del material | Tipo | Enlace |
| Análisis de datos | Aguilar, J. (2021). *Estadística descriptiva, regresión y probabilidad con aplicaciones.* Ediciones de la U. | Libro digital | <https://www-ebooks7-24-com.bdigital.sena.edu.co/?il=15156&pg=202> |
| Análisis de datos | Moreno, A. & Córcoles, S. (2019). *Python* *práctico.* Editorial Rama. | Libro digital | <https://www-ebooks7-24-com.bdigital.sena.edu.co/?il=10285&pg=1> |
| Modelos estadísticos | Triola, M. (2018). *Estadística* (12 edición). Pearson Educación. | Libro digital | <https://www-ebooks7-24-com.bdigital.sena.edu.co/?il=7386&pg=149> |
| Modelos estadísticos | Toomey, D. (2018). *Learning Jupyter 5: Explore interactive computing u*  *sing Python, Java, JavaScript, R, Julia, and JupyterLab* (2nd Edition). Packt Publishing. | Libro digital | <https://search-ebscohost-com.bdigital.sena.edu.co/login.aspx?direct=true&db=e000xww&AN=1881491&lang=es&site=ehost-live> |
| Modelos estadísticos | Toomey, D. (2017) *Jupyter for data science.* Packt Publishing. | Libro digital | <https://search-ebscohost-com.bdigital.sena.edu.co/login.aspx?direct=true&db=e000xww&AN=1637911&lang=es&site=ehost-live>. |

**Glosario**

| **Tipo de recurso** | Glosario |
| --- | --- |
| Censo: | conjunto de actividades destinadas a recoger, recopilar, el padrón o lista de la población de una nación o pueblo en un momento determinado, la información demográfica social y económica, para luego evaluar, analizar y publicar o divulgar. |
| Continua: | una variable continua es aquella que puede tomar un número infinito de valores entre dos valores cualesquiera de una característica. la altura de 5 personas: 1.85, 1.89, 1.93, 1.59, 1.65. En la práctica medimos la altura con dos decimales, pero también se podría dar con tres decimales o más. |
| Cualitativa: | variable en la que los resultados posibles no son valores numéricos. Por ejemplo: color del pelo, tipo de ropa preferida, lugar de veraneo, etc. |
| Cuantitativa: | se trata de características cuyas variaciones pueden ser expresadas de forma numérica. Por ejemplo: edad, peso, estatura, número de hijos, años de servicio, etc. |
| Cuartiles: | son 4 divide un conjunto de datos ordenados en cuatro partes iguales. Cuartil 1 (Q1) 25 %, la mediana es el Cuartil 2 (Q2) 50 %, Cuartil 3 (Q3) 75 %, Cuartil 4 (Q4) 100 %. Se trabaja principalmente con los Q1, Q2, Q3. |
| Dato atípico: | es un valor mayor a 1.5 veces el rango intercuartil. |
| Desviación estándar: | mide la dispersión de una distribución de datos. Entre más dispersa está una distribución de datos, más grande es su desviación estándar. |
| Discreta: | Una variable es discreta cuando no puede tomar ningún valor entre dos consecutivos, y que es continua cuando puede tomar cualquier valor dentro de un intervalo. Ejemplos de variable discreta: número de empleados de una fábrica; número de hijos; número de cuentas en un banco. |
| Escala cronológica: | el orden cronológico de los sucesos es la ubicación de éstos en una línea temporal. |
| Escala nominal: | es una escala de medición en la cual los números sirven como “etiquetas” solamente para identificar o clasificar un objeto. Una escala de medición nominal normalmente trata sólo con variables no numéricas (no cuantitativas). |
| Escala ordinal: | permite evaluar la actitud de un encuestado con respecto a un tema utilizando un conjunto de respuestas ordenadas. Por ejemplo, las respuestas pueden incluir: “muy satisfecho”, “satisfecho”, “insatisfecho” y “muy insatisfecho”. |
| Estadísticos: | es cualquier función real medible de la muestra de una variable aleatoria. |
| Media: | resultado que se obtiene al dividir la suma de varias cantidades por el número de sumandos. |
| Muestreo: | selección de un conjunto de personas o cosas que se consideran representativos del grupo al que pertenecen, con la finalidad de estudiar o determinar las características del grupo. |
| Rango intercuartil: | es una medida de dispersión. Se obtiene restando el cuartil 3 (Q3) menos el cuartil 1(Q1); Q3-Q1 en ese rango está el 50 % de los datos. |

**Referentes bibliográficos**

| **Tipo de recurso** | Bibliografía |
| --- | --- |
| Alcaldía de Medellín. (2022). *MEData*. <http://medata.gov.co/dataset/> | |
| Altair. (2022). *Vega-Altair: visualización declarativa en Python.* <https://altair-viz.github.io/> | |
| Carto. (2022). *Análisis espacial para la pila de datos moderna.* https://carto.com/ | |
| Lind, D., Whaten, S. & Marchal, W. (2019). *Estadística aplicada a los negocios y la economía* (17 edición). McGraw Hill. <https://www-ebooks7-24-com.bdigital.sena.edu.co/?il=9246&pg=315> | |
| Martínez, C. (2019). *Estadística y muestreo* (14 edición). Ecoe ediciones. <https://www-ebooks7-24-com.bdigital.sena.edu.co/?il=9278&pg=387> | |
| Martínez, C. (2019). *Estadística básica aplicada* (5ta edición). Ecoe ediciones. <https://www-ebooks7-24-com.bdigital.sena.edu.co/?il=8975&pg=82> | |
| Matplotlib. (2022). *Visualización con Python.* <https://matplotlib.org/> | |
| Microsoft. (2022). *Power BI.* <https://powerbi.microsoft.com/es-es/> | |
| Plotly. (2022). *Biblioteca de gráficos de código abierto para Python*. <https://plotly.com/python/> | |
| Seaborn. (2022). *Visualización de datos estadísticos*. <https://seaborn.pydata.org/> | |
| Tableau Public. (2022). *Tableau Public*. <https://public.tableau.com/app/discover> | |
| Vladimirovna, O. & Gutiérrez, E. (2016). *Estadística inferencial 1 para ingeniería y ciencias*. Grupo Editorial Patria.  <https://elibro-net.bdigital.sena.edu.co/es/ereader/senavirtual/40474?page=69> | |