Guía 1 – Conociendo Orange Data Mining

El objetivo de esta guía es familiarizarse con Orange Data Mining. En esta guía realizaremos algunas labores de preprocesamiento y visualización de datos mediante flujos de trabajo visual.

¿Qué es Orange Mining?

Orange es una herramienta de minería de datos y visualización de datos de código abierto que ofrece una amplia gama de características. Algunas de las características destacadas de Orange Mining son:

Interfaz Gráfica Intuitiva

Orange proporciona una interfaz gráfica de usuario (GUI) intuitiva y fácil de usar que permite a los usuarios realizar análisis de datos sin necesidad de programación.

Flujos de Trabajo Visuales

Los usuarios pueden construir flujos de trabajo visuales arrastrando y soltando widgets (bloques de construcción de análisis de datos) en el lienzo y conectándolos para crear procesos de análisis complejos.

Amplia Gama de Widgets

Orange ofrece una amplia gama de widgets para realizar diversas tareas de análisis de datos, incluyendo carga de datos, preprocesamiento, modelado, visualización y evaluación de modelos.

Preprocesamiento de Datos

Los usuarios pueden realizar diversas operaciones de preprocesamiento de datos, como limpieza de datos, selección de características, transformación de datos, y más, utilizando widgets específicos.

Modelado de Datos

Orange permite construir y evaluar modelos de aprendizaje automático utilizando una variedad de algoritmos, incluyendo clasificación, regresión, clustering, asociación y detección de anomalías.

Visualización de Datos

Los usuarios pueden visualizar sus datos de diversas formas utilizando widgets de visualización, incluyendo gráficos de dispersión, histogramas, diagramas de caja, redes, mapas de calor, y más.

Evaluación de Modelos

Orange proporciona herramientas para evaluar y comparar el rendimiento de los modelos de aprendizaje automático utilizando técnicas de validación cruzada, curvas de aprendizaje, matrices de confusión, y más.

Extensibilidad y Personalización

Orange es altamente extensible y permite a los usuarios crear sus propios widgets personalizados o instalar widgets adicionales desarrollados por la comunidad de usuarios de Orange.

Instrucciones

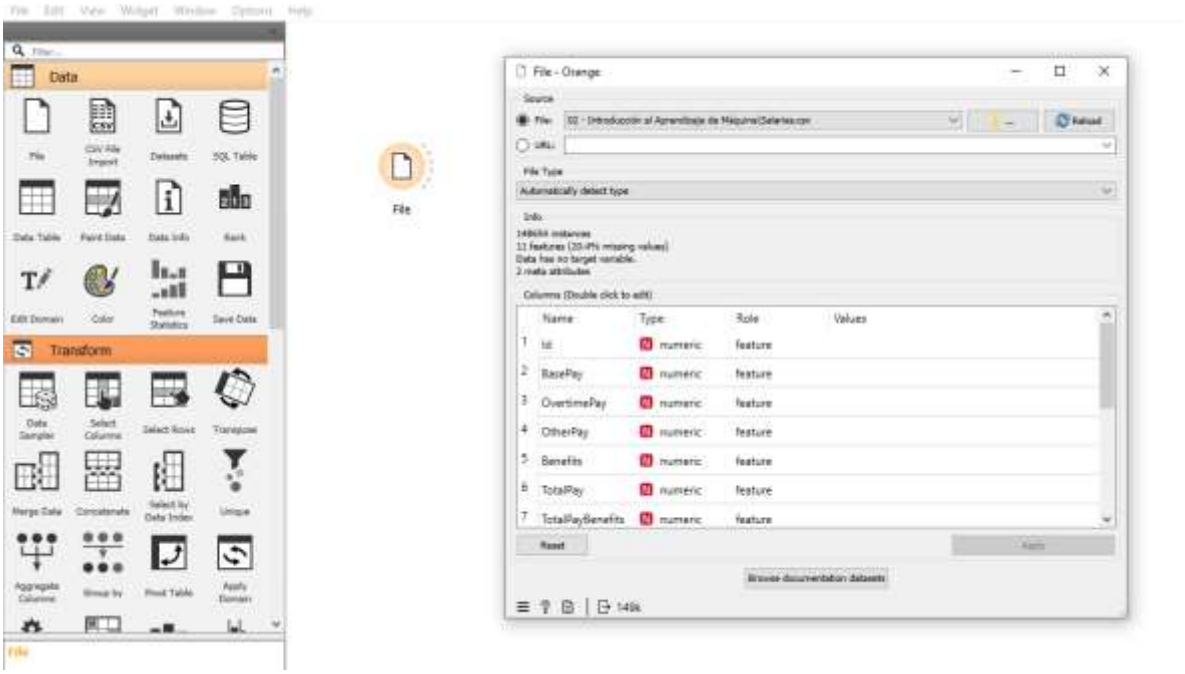
Vaya siguiendo paso a paso las indicaciones para realizar las tareas solicitadas. Deberá guardar su archivo con los flujos creados para posteriormente subirlo a su carpeta de tareas y entregables.

Análisis Sueldos San Francisco

A continuación, realizaremos un análisis de los sueldos de San Francisco a partir del set de datos utilizado en módulos anteriores.

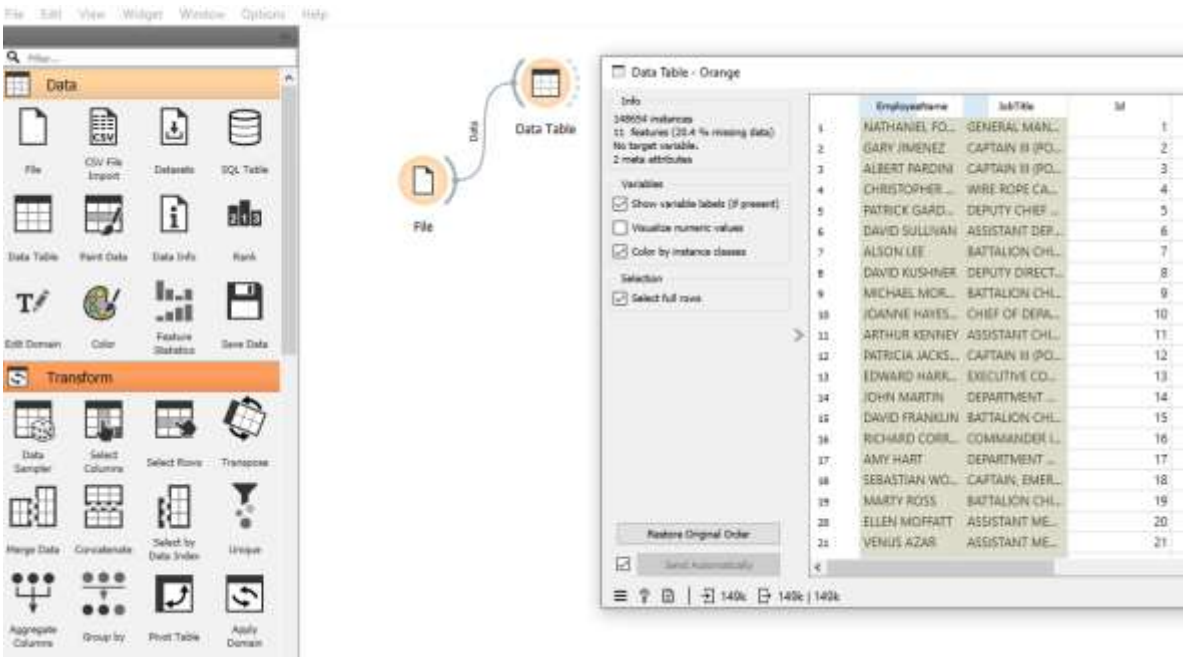
1. **Lectura de Datos**

Cargue el set de datos Salaries.csv en un componente de tipo File.



2. **Visualización inicial**

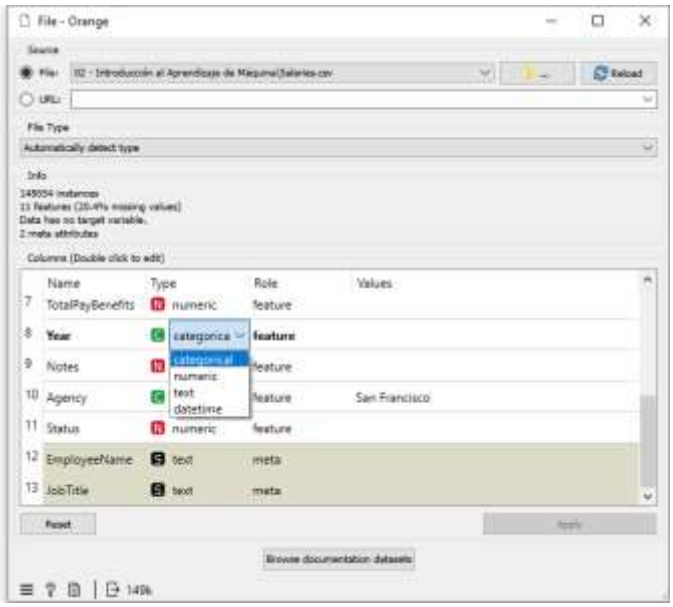
Conecte un widget de tipo Tabla para desplegar el contenido del set de datos.

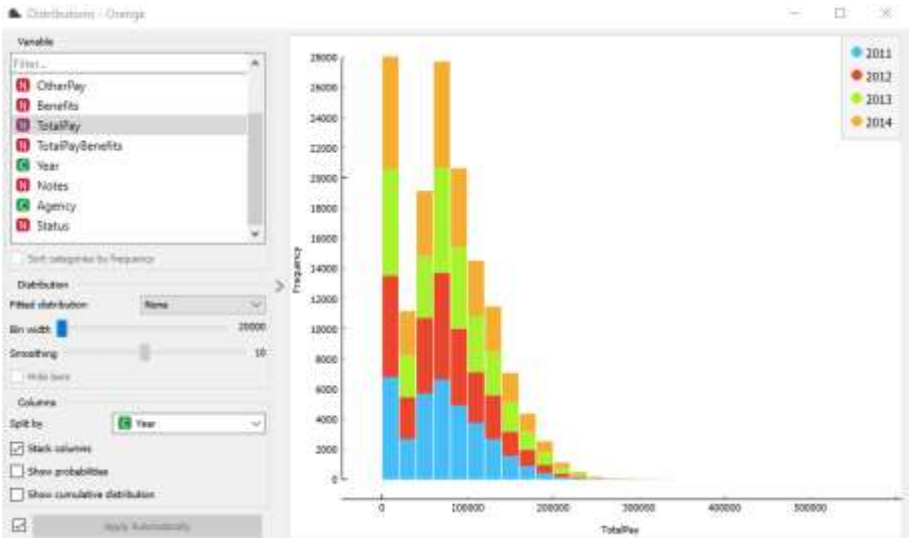


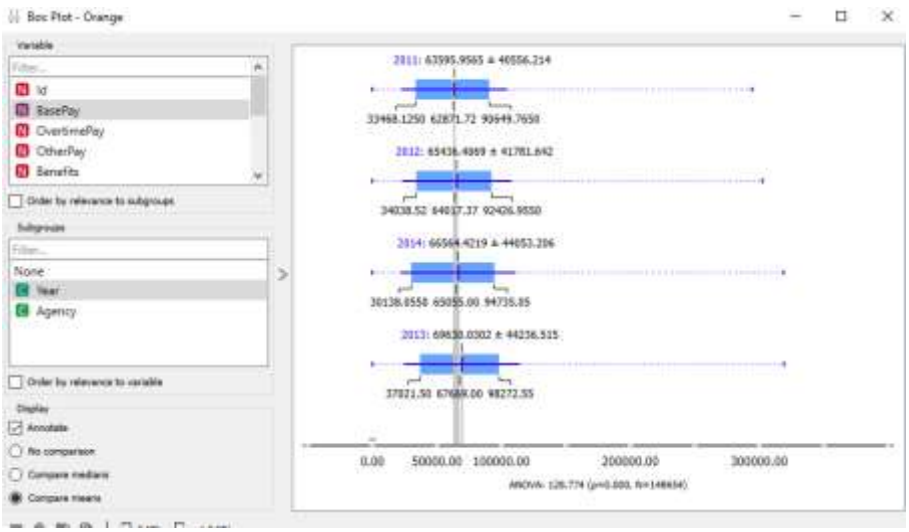
3. Análisis de la columna BasePay

Realice un análisis univariado de la columna BasePay utilizando los elementos de la paleta “Visualize”. Utilice lo aprendido en el módulo de análisis exploratorio de datos. Al menos, utilice un histograma y un diagrama de caja y bigote para caracterizar la variable.

Note que, si define la columna Year como categórica, puede utilizarla en las visualizaciones para separar las series de datos.


File - Orange
Source
File:
URL:
Reload
File Type
Automatically detect type
Info
Columns (Double click to edit)
Name
Type
Role
Values
7 TotalPayBenefits numeric feature
8 Year categorical feature
9 Notes numeric feature
10 Agency text feature San Francisco
11 Status datetime feature
12 EmployeeName text meta
13 JobTitle text meta
Reset
Apply
Browse documentation datasets

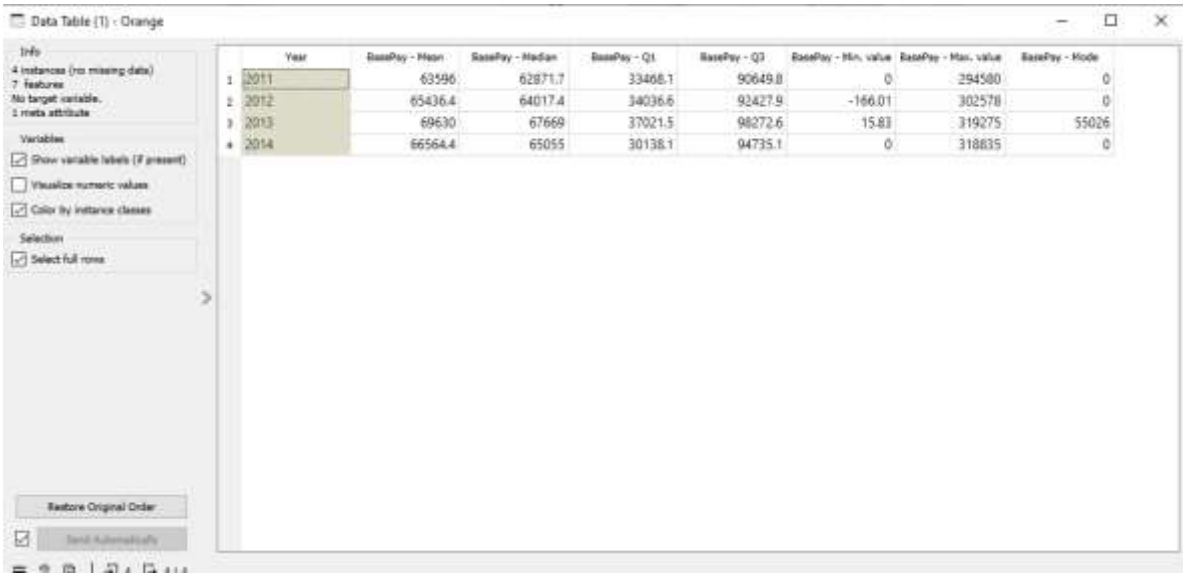

Variable
OtherPay
Benefits
TotalPay
TotalPayBenefits
Year
Notes
Agency
Status
Distribution
Fitted distribution
None
Bin width
Smoothing
Columns
Split by
Year
Stack columns
Show probabilities
Show cumulative distribution
2011
2012
2013
2014
Frequency
TotalPay



4. Agrupamiento de datos

Utilice el widget de agrupamiento de datos para obtener distintas agregaciones de la columna BasePay, agrupe por año. Puede revisar la documentación del widget y mirar la imagen de ejemplo.

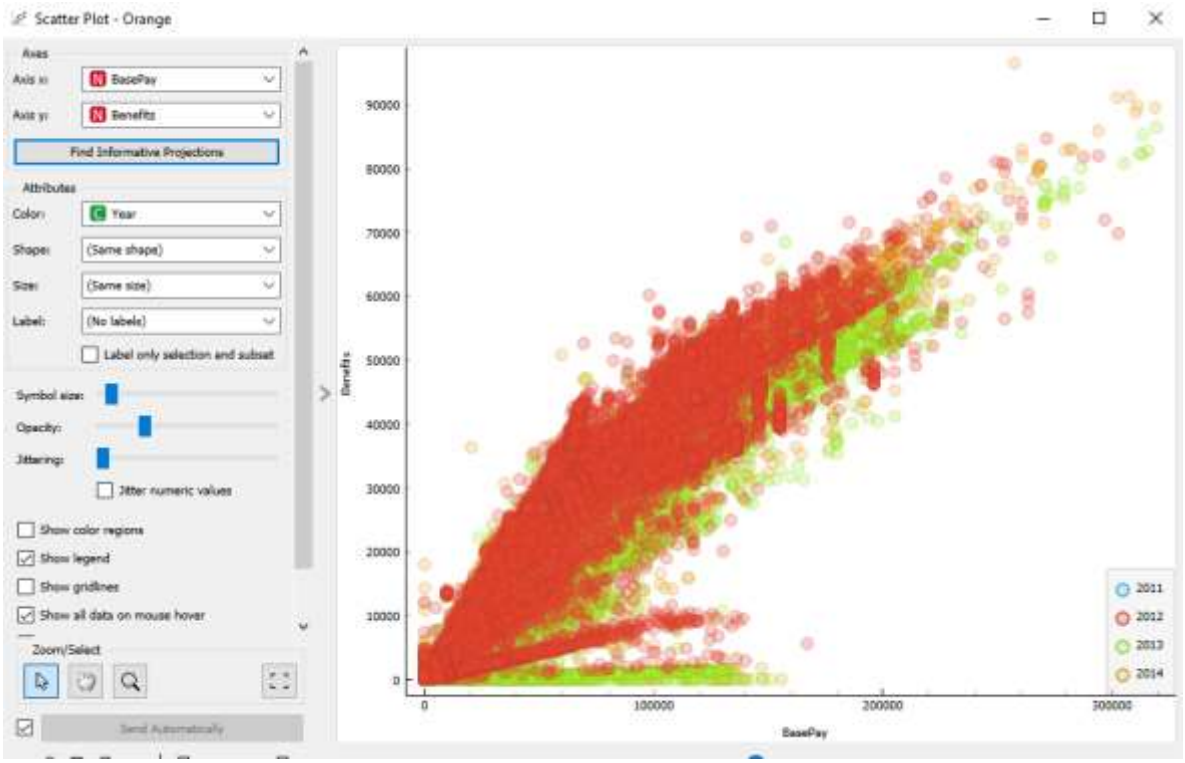
https://orangedatamining.com/widget-catalog/transform/groupby/



	Year	BasePay - Mean	BasePay - Median	BasePay - Q1	BasePay - Q3	BasePay - Min. value	BasePay - Max. value	BasePay - Mode
1	2011	63596	62871.7	33468.1	90649.8	0	294580	0
2	2012	65436.4	64017.4	34036.6	92427.9	-166.01	302578	0
3	2013	69630	67669	37021.5	98272.6	15.83	319275	55026
4	2014	66564.4	65055	30138.1	94735.1	0	318835	0

5. Relación entre BasePay y Benefits

Haga un análisis visual donde se aprecie la correlación entre BasePay y Benefits.

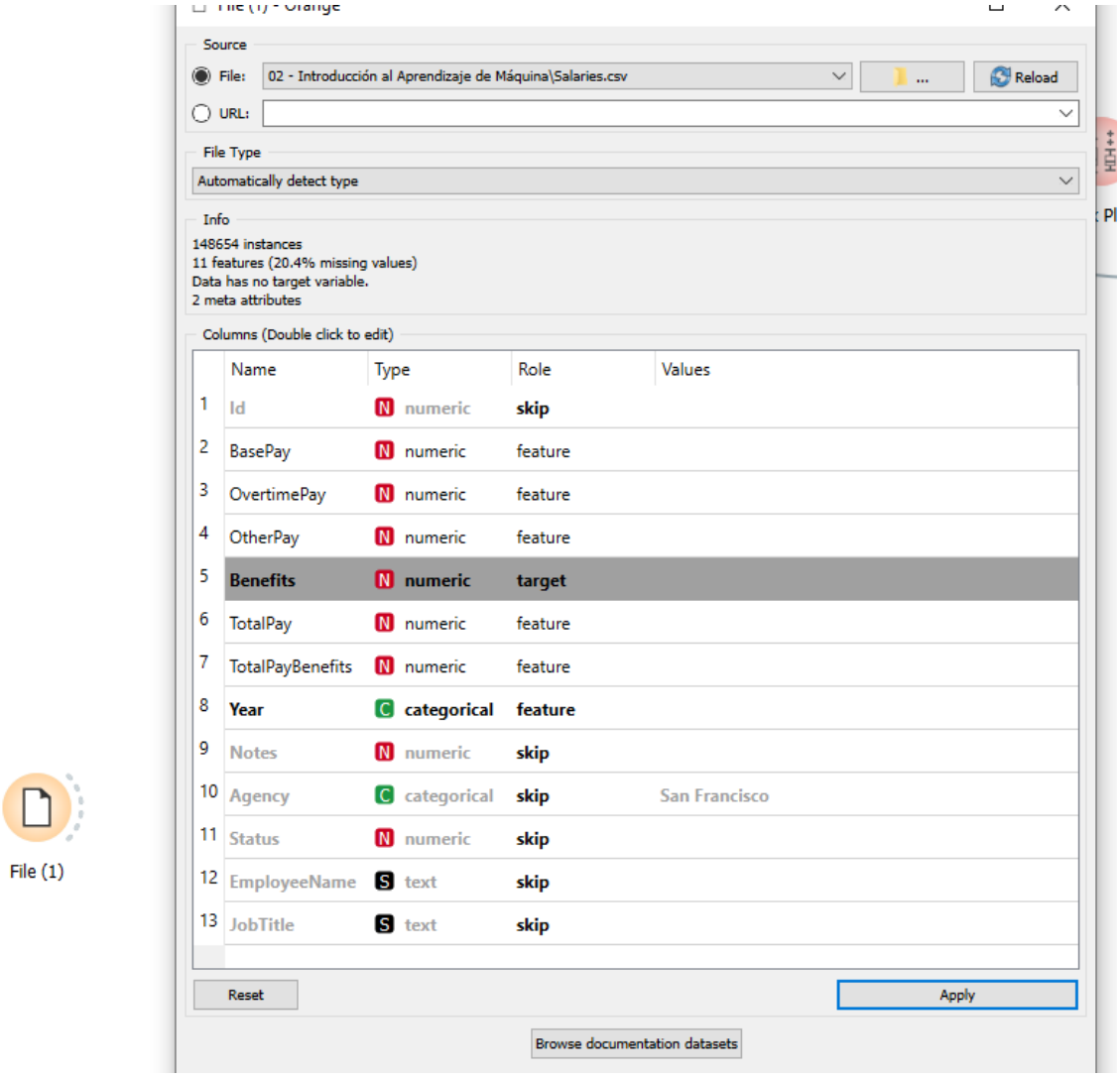


6. Analice los datos perdidos en la columna BasePay

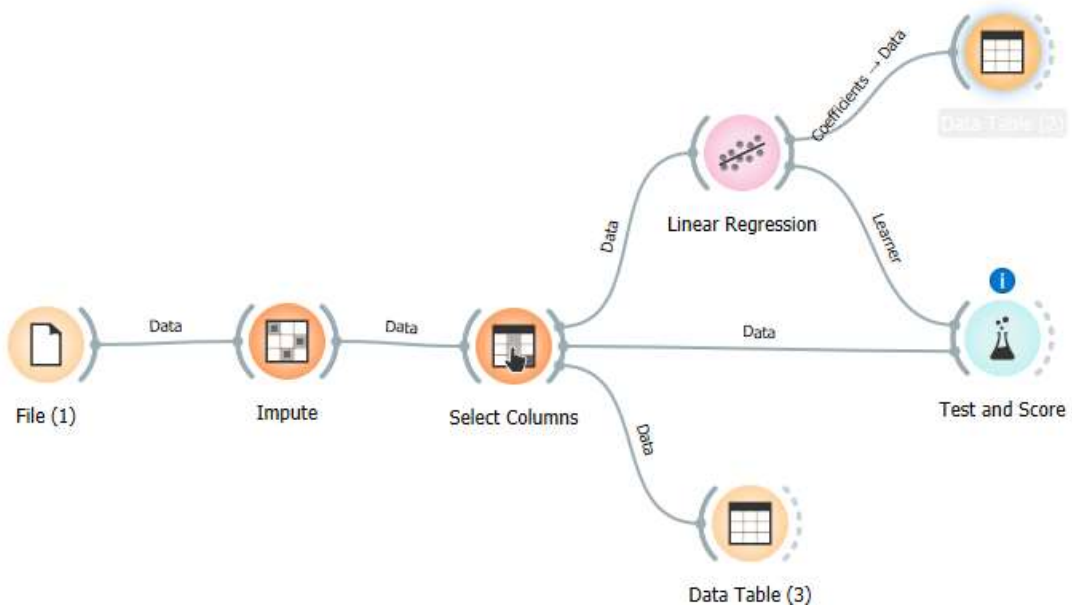
¿Cómo se puede analizar los valores perdidos de la columna BasePay? No hay ningún widget para este efecto, pero tal vez una combinación de widgets podría ayudar.

7. Modelo regresivo simple BasePay /

Ahora vamos a plantear un modelo regresivo, pero primero, sugerimos clonar el widget File que carga los datos y realizar una definición de variables como la siguiente (recuerde presionar Apply). En este modelo, vamos a definir la variable Benefits como la variable de outcome (dependiente), la variable BasePay será la variable predictora. Es decir, un modelo en donde suponemos que el sueldo base explica los beneficios.



Para implementar este flujo, se recomienda utilizar los siguientes widgets:



El widget **Impute** se hace cargo de tratar los valores nulos, en este caso, se ha optado por eliminar filas con valores nulos.

El widget **Select Columns**, permite seleccionar los predictores que se agregarán al modelo. En este caso, solamente se ha incorporado la variable independiente BasePay. La tabla conectada, muestra el set de datos después de realizar la selección de columnas.

El widget **Linear Regression** realiza el ajuste del modelo, y la tabla conectada a la salida, despliega los coeficientes de la regresión.

El widget **Test & Score**, despliega las métricas de evaluación del modelo regresivo.

¿Qué resultados obtiene en la evaluación de este modelo regresivo?