

UNIVERSIDAD NACIONAL DE COLOMBIA – SEDE MANIZALES

FACULTAD DE INGENIERÍA Y ARQUITECTURA

DEPARTAMENTO DE INGENIERÍA ELÉCTRICA Y ELECTRÓNICA

ANALÍTICA DE DATOS

INFORME Nº1

TALLER PRÁCTICO DE ORANGE 1

PRESENTADO A: PROF. DIEGO COLLAZOS Y

SANTIAGO PINEDA QUINTERO

PRESENTADO POR: FELIPE FERNÁNDEZ ALZATE (1056120378)

PRIMERA PARTE: ANÁLISIS DE 2 BASES DE DATOS ALEATORIAS

La primera base de datos que vamos a analizar se trata de “Cáncer de pulmón por consumo de cigarrillo”, la cual tiene 309 instancias (personas) y 16 atributos listados a continuación:

1. Género 9. Si tiene alergias
2. Edad 10. Si jadea constantemente
3. Si es fumador 11. Si consume alcohol
4. Si tiene dedos amarillos 12. Si tose
5. Si sufre de ansiedad 13. Si respira con dificultad
6. Si sufre de presión grupal 14. Si traga con dificultad
7. Si tiene enfermedades crónicas 15. Si tiene dolor en el pecho
8. Si tiene fatiga 16. Si sufre cáncer de pulmón

Para la mayorías de los atributos se manejan los valores 1 y 2, siendo, respectivamente, un No y un Si.

En base a esta información podemos generar diversas relaciones importantes para obtener información relevante acerca de los fumadores.

En el primer ejemplo, vemos la distribución de edad entre los encuestados, según el siguiente diagrama:

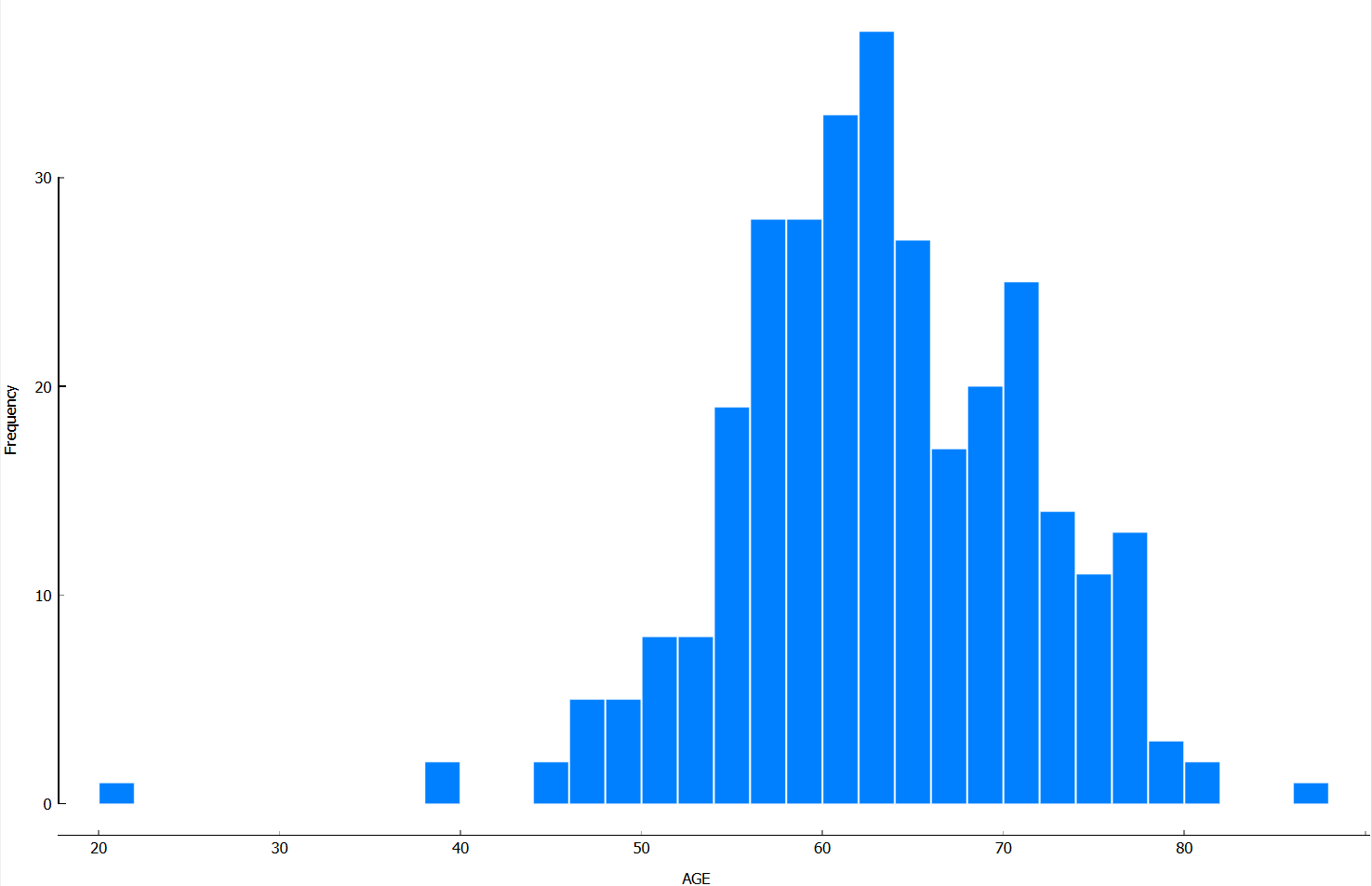


Ilustración 1.Distribución de edad en la encuesta

Se puede observar que la mayoría de encuestados se encuentra en el rango de 55 a 57 años de edad, siendo el mayor porcentaje el rango de [62,64] años con 11.97% del total, mientras que los menores rangos (con valores) son los menores a 22 y mayores a 86, con solo 1 representante que equivale al 0.32%. Si se toma el mismo diagrama y habilitamos una distribución normal, se podrá observar en los datos dados por Orange que la media de edad en la encuesta es µ=62.67 años:

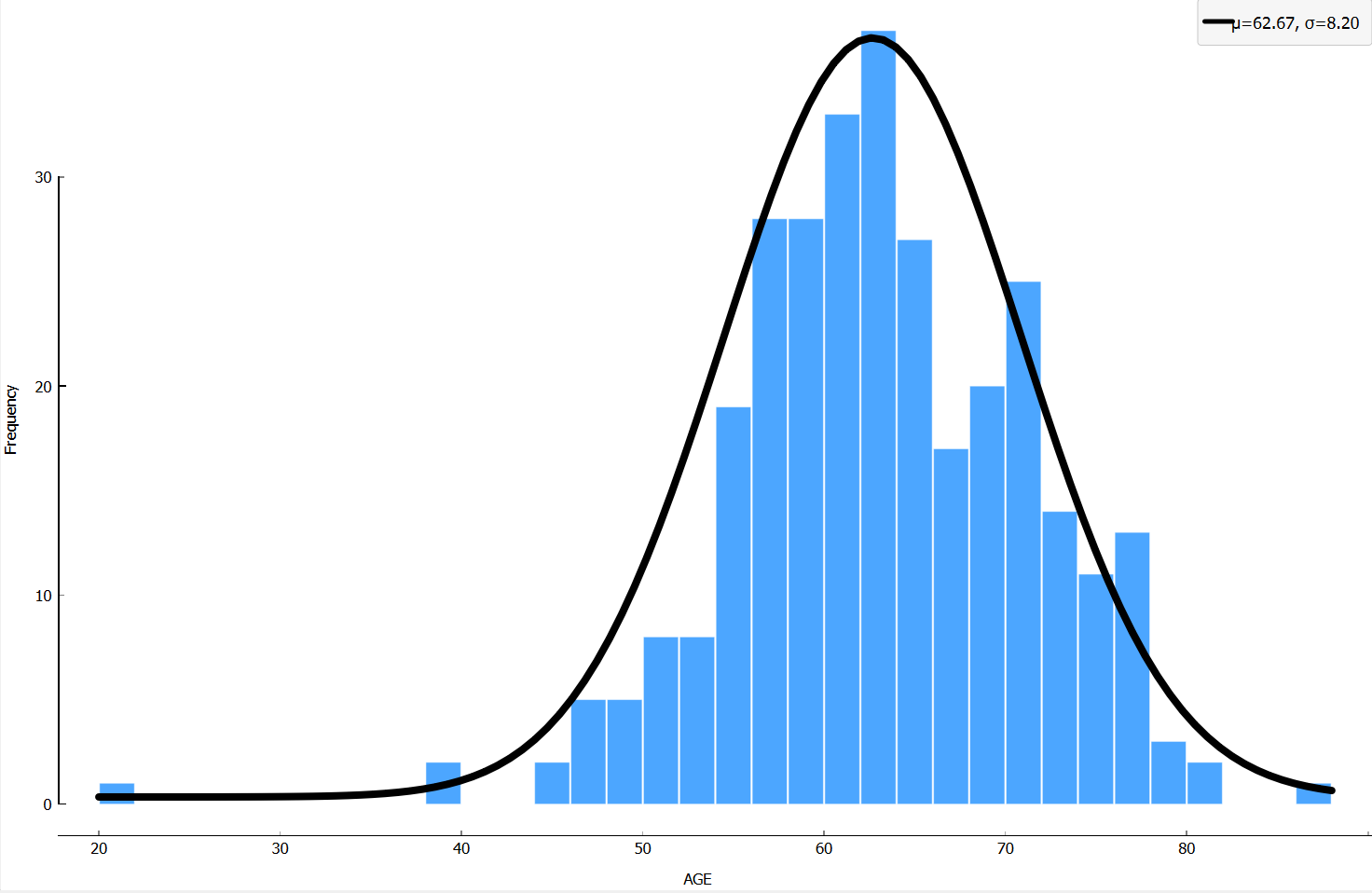


Ilustración 2. Diagrama con Distribución normal aplicada

En el segundo ejemplo, se genera un diagrama de Venn para ver la similitud entre las variables de “Fumadores” , “Presión grupal” y “Alcohólicos”, como se ve a continuación:

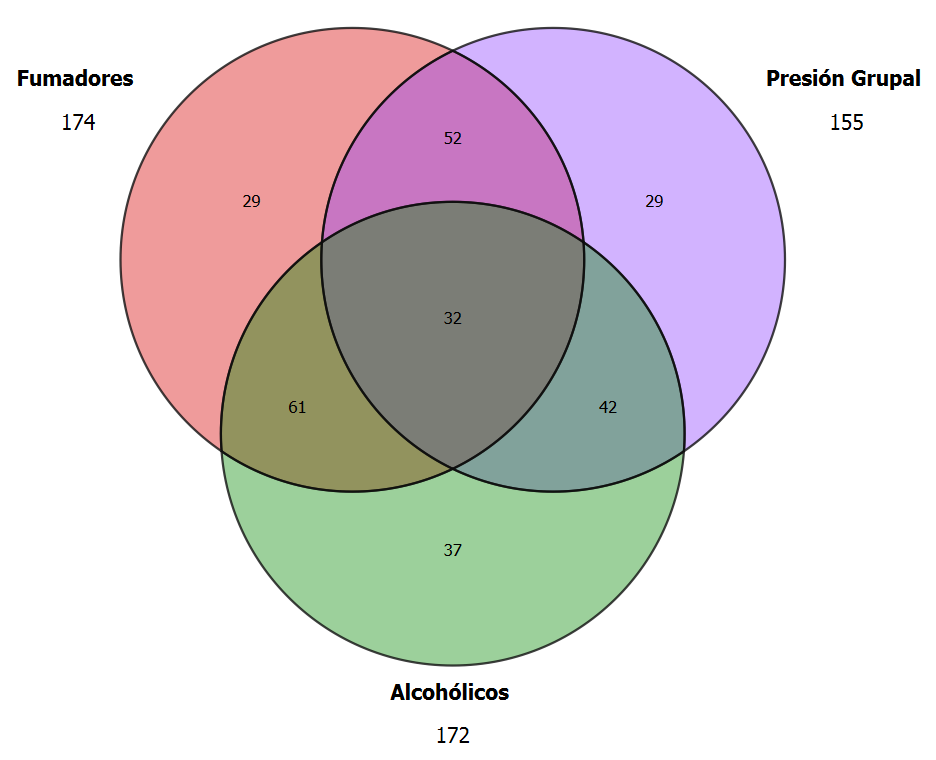


Ilustración 3.Diagrama de Venn entre Fumadores, Presión grupal y alcohólicos

Según el diagrama podemos concluir varias cosas:

* El total de fumadores en la encuesta es de 174 (56.31%)
* El total de personas que sufren de presión grupal en la encuesta es 155 (50.16%)
* El total de alcohólicos en la encuesta es de 172 (55.66%)
* Las personas que fuman y sufren de presión grupal son 84 (27.18%)
* Las personas que fuman y beben alcohol es 61 (19.74%)
* Las personas que beben alcohol y sufren presión grupal son 42 (13.59%)
* Las personas que fuman, beben alcohol y sufren de presión grupal son 32 (10.35%)

Con este análisis del diagrama de Venn se podría pensar que la presión grupal es un componente importante a tener en cuenta para que una persona sea fumadora o alcohólica; también se puede ver que gran parte de las personas encuestadas tienen el hábito de fumar y beber alcohol.

Por último, vamos a relacionar la edad de los encuestados y si sienten algún dolor en el pecho con si sufren de cáncer de pulmón, para entendimiento del diagrama, los círculos indican que no sufren de dolor en el pecho y las “X” que sí; las 2 columnas por parte del eje X es para determinar si tienen o no cáncer de pulmón; y en el eje Y se encuentran los rangos de edad de los encuestados.

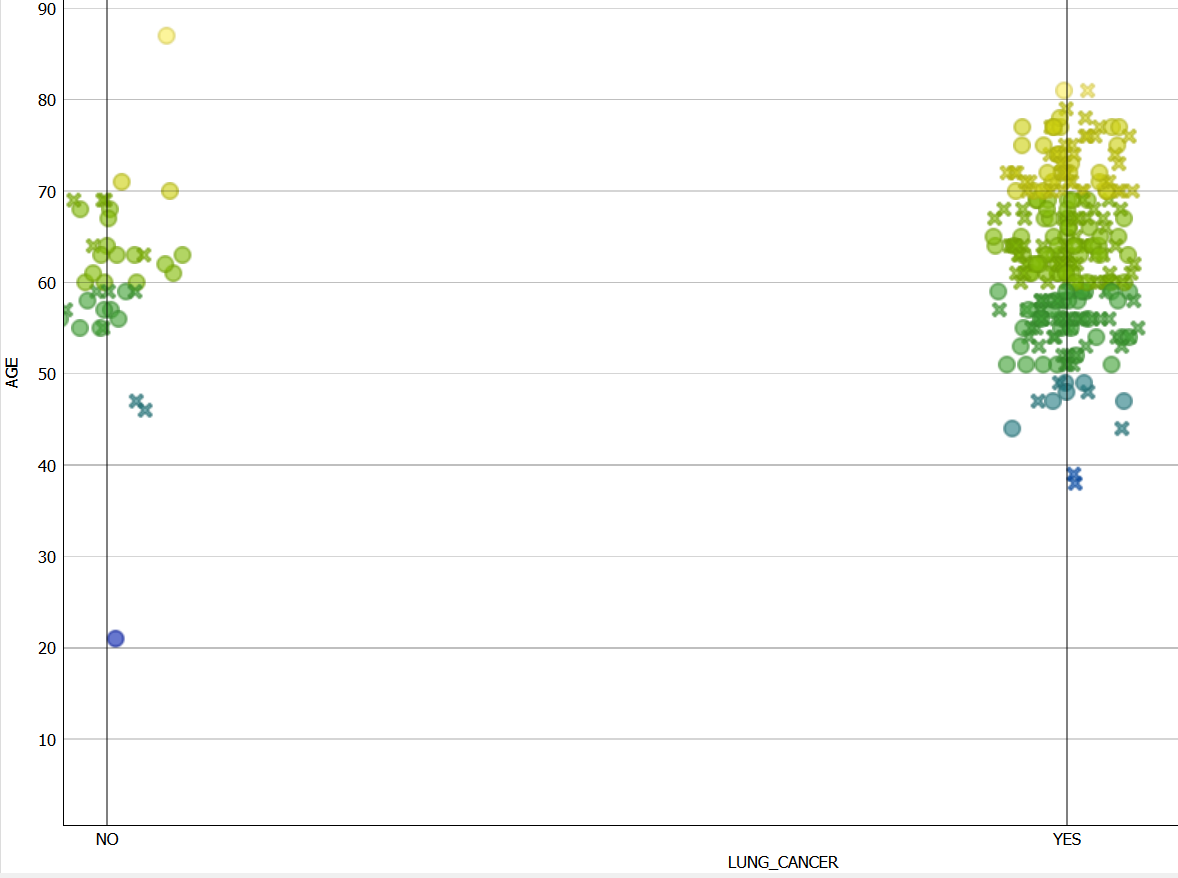


Ilustración 4.Scatter plot que muestra similitudes entre Cáncer de pulmón, edad y dolor en el pecho

Gracias a este gráfico se puede observar que gran parte de las personas que sufren de cáncer de pulmón presentan dolores en el pecho constantes, al igual que se puede analizar que el rango de edad con mayor riesgo de contraer este tipo de cáncer es entre los 40 y los 80 años.

Conclusiones:

* En el caso de una tienda, gracias a los datos visualizados y la información obtenida, se puede pensar en una publicidad donde se ofrezcan bebidas alcohólicas a la par de cajetillas de cigarrillos para que ese porcentaje de gente que entra en ambas categorías gaste mas de su dinero allí.
* En el caso de un programa de salud, hay opciones de crear campañas para evitar el cáncer de pulmón en personas con una edad cercana a los 62 años (o desde una edad temprana-25 años) y dar a conocer los síntomas relacionados a éste.
* En el caso de un programa social de acompañamiento juvenil, se pueden realizar charlas para evitar que las personas ingresen al mundo del consumo de bebidas alcohólicas y cigarrillo por culpa de la presión de personas exteriores o de un grupo social cercano.

El diagrama completo para la visualización de los datos en Orange quedó de la siguiente manera: una tabla que conecta con 1 distribución, 3 selecciones de filas que ingresan al diagrama de Venn y un Scatter plot:

Se puede ingresar al archivo en el siguiente enlace: [CáncerDePulmón](https://drive.google.com/file/d/1Teb5XdABVkiHRTxoProIZ30JwJZ0_iZk/view?usp=sharing)

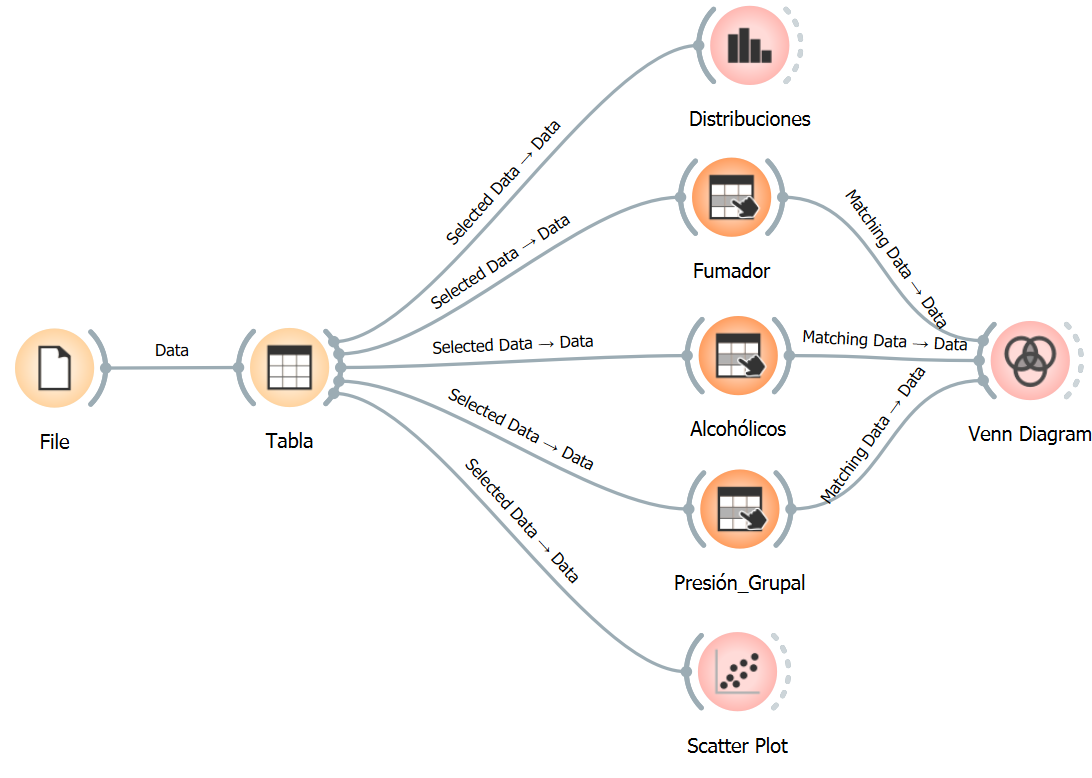


Ilustración 5. Flujo de trabajo construido en Orange

La segunda base de datos que vamos a trabajar trata de “Precios de la gasolina en estados de Canadá”, en la cual se verá el comportamiento de los precios base y con impuestos a lo largo de un periodo de 6 años (2018-Actualidad). Cuenta con 219 instancias y 15 categorías:

1. Fecha 9.Montreal
2. Toronto 10.Winnipeg
3. Ottawa 11.Regina
4. Thunder Bay 12.Calgary
5. St.John Newfoundland 13.Vancouver
6. Charlottetown 14.Estado de impuesto
7. Halifax 15.Situación fiscal
8. St.John New Brunswick

El primer ejercicio hecho con esta base de datos fue realizar un scatter plot con varios estados y la variación de los precios en los mismos; en general, las gráficas se ven así:

A graph of different colored dots

Description automatically generated

Ilustración 6. Scatter plot con la variación del precio de la gasolina en Montreal

En la ilustración anterior se ve el comportamiento de los precios de la gasolina en el estado de Montreal. Lo que mas se puede comprender en el diagrama es que los precios de los impuestos se mantienen bastante estables, mientras que los precios base fluctúan mucho con el paso del tiempo. Tienen su punto máximo en el primer semestre del 2022 y el punto mínimo a inicios del 2020. Así funciona en cada scatter plot que se realiza con otros estados.

Otro análisis que se puede realizar va de la mano con una distribución del precio total de la gasolina; para este caso, se va a analizar la ciudad de Ottawa:

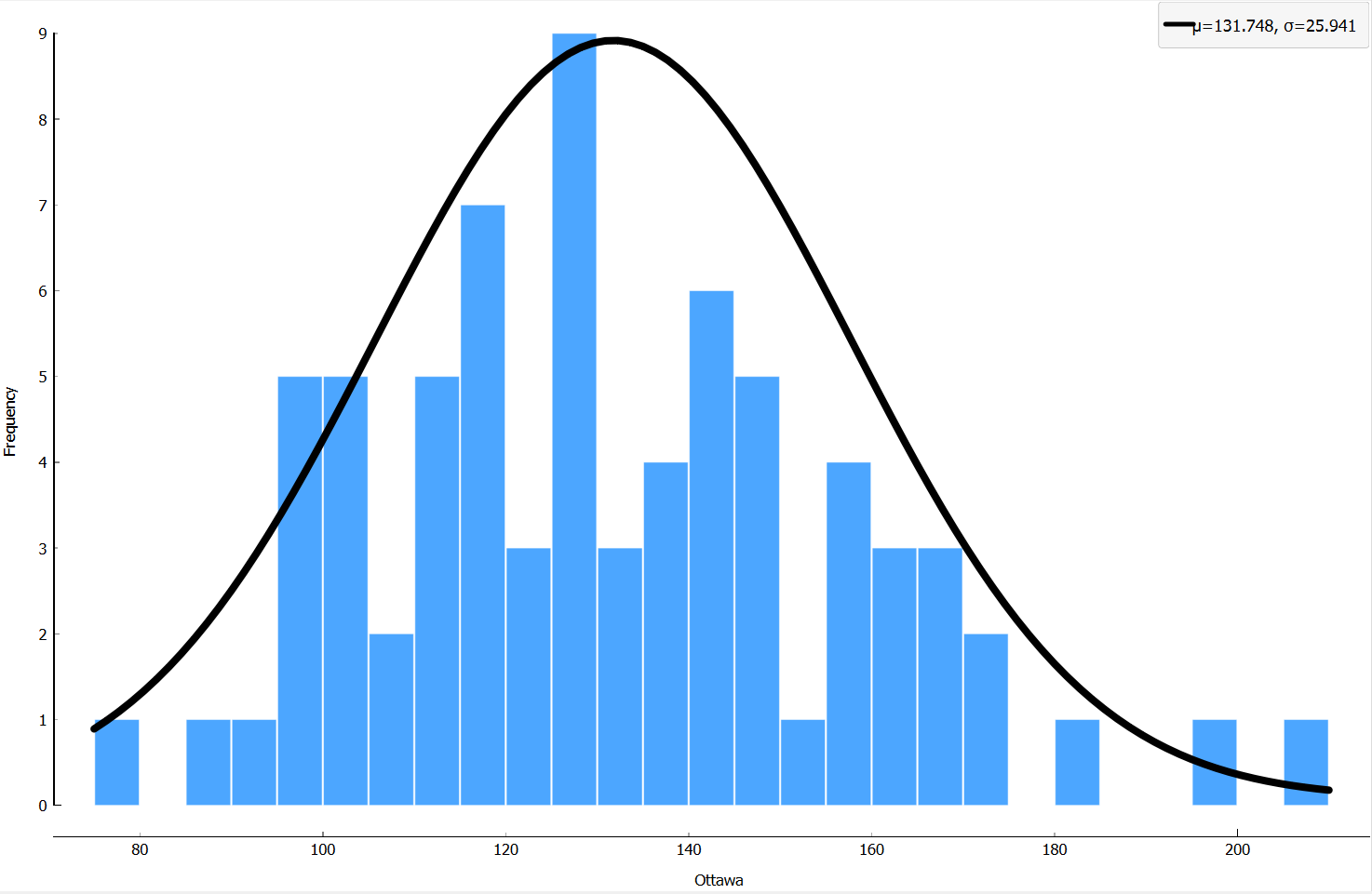


Ilustración 7. Distribución de los precios totales a lo largo del tiempo en Ottawa

Gracias a este gráfico podemos concluir que la mayoría de los precios se encuentran entre 120 y 140, pero, aún así, la media del precio de la gasolina es 131.748 centavos por litro. Realizando el mismo proceso con otros estados se obtienen los siguientes resultados para la media del precio:

* Toronto: 133.33 c/l
* Thunder Bay: 142.167 c/l
* Winnipeg: 128.170 c/l
* Calgary: 124 c/l
* Vancouver: 161.225 c/l

Estos datos sirven para realizar un estudio de la diferencia de precios en diferentes estados de Canadá, teniendo en cuenta el estatus social de las personas que allí residen, sus ingresos y el impuesto sobre la gasolina.

En el último análisis hecho, se tomó una pequeña muestra aleatoria de 5 fechas para observar el valor mínimo y valor máximo en todos los estados presentes en la base de datos, esto con el fin de realizar comparaciones entre cada estado en momentos aleatorios:

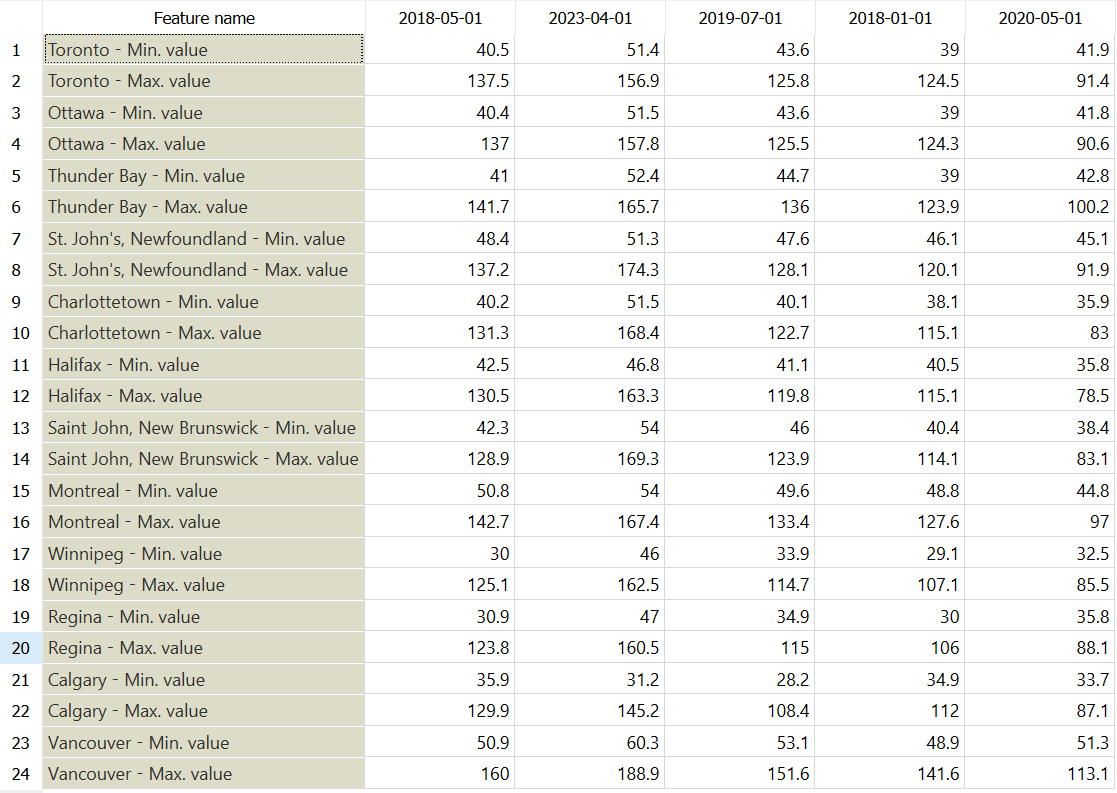
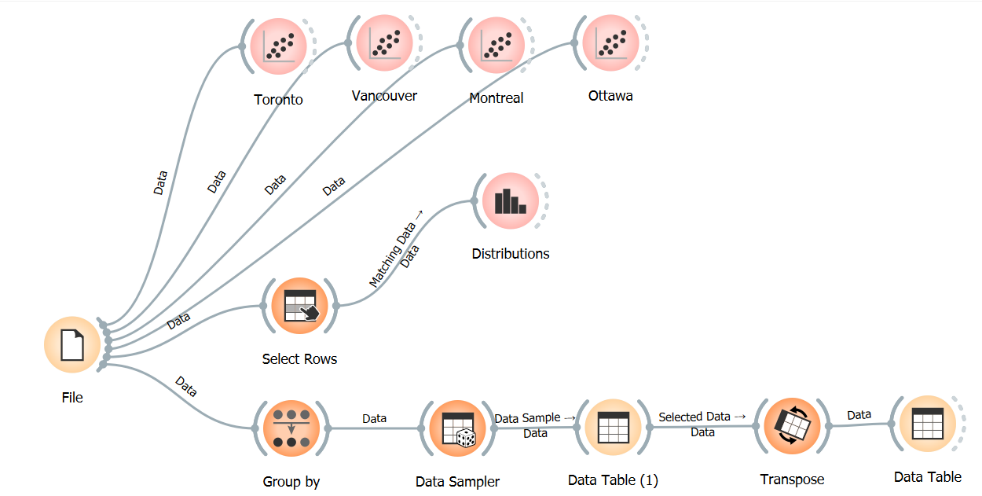


Ilustración 8. 5 fechas aleatorias y valores mínimos y máximos de gasolina

Para comprender esta tabla, se tomará a Regina como ejemplo; para el 1º de Enero de 2018, el valor mínimo de la gasolina era de 30 cents/litro (que equivale al impuesto mismo), y llegaba a un máximo de 106 cents/litro, gracias a esto se puede realizar una estimación del valor base de la gasolina en ese momento.

Luego podemos analizar el 1º de Mayo de 2018, el valor mínimo era de 30.9 cents/litro y el máximo 123.8 cents/litro. El valor de los impuestos no sufre cambios drásticos en el transcurso de 4 meses, sin embargo, se nota que el precio base (que puede ser estimado) si tiene cambios bruscos, lo cual nos permitiría hacer un análisis del mercado y de por qué este precio varía tanto.

El enlace para el archivo a continuación: [Gasolina](https://drive.google.com/file/d/1OO1unUDXNNQaueICzj1wylsvt7LSQasK/view?usp=sharing)



SEGUNDA PARTE: COMPRENSIÓN DE LOS WIDGETS EN ORANGE

1. **Correlation\_example**

Este archivo cuenta con 5 widgets:

Datasets: Carga un repositorio en línea de una lista selecta de bases de datos en diferentes idiomas, en este caso, una base de datos de la calidad de vida en diversos países del mundo.

Data Table: En un cuadro/tabla, muestra toda la base de datos, con todas sus instancias, categorías e incluso brinda información del porcentaje de datos faltantes en la base.

Correlations: En base a la correlación de Pearson, la cual busca la similitud lineal entre 2 variables, obtenemos una lista de todas las posibles correlaciones entre las variables de nuestra base de datos, nos indica qué variables se tomaron y cual es su coeficiente de correlación. Este coeficiente varía entre -1 y 1 así: -1 indica una correlación negativa (mientras una aumenta, la otra disminuye o viceversa), 0 indica que no hay correlación (las variables no se relacionan en nada) y 1 indica una correlación positiva (ambas variables aumentan o disminuyen equivalentemente)

Scatter Plot: Toma 2 variables para cada uno de los ejes, y en una representacion de 2 dimensiones ayuda a analizar el comportamiento, patrones o relaciones de las variables, cada instancia representa un punto en el mapa.

Data Table (1): En base a los resultados de Correlations, genera una tabla con todas las relaciones de las variables y su respectiva correlación.

1. **Feature\_selection\_example**

Para evitar repeticiones innecesarias se tendrán 6 widgets en esta base de datos, obviando Data Table:

Datasets: Del mismo modo que anteriormente, carga un repositorio en línea, en este caso, acerca de las carácterísticas de un conjunto de vinos.

Distributions y Distributions(1): En base a una o varias categorías seleccionadas, pinta una gráfica de la distribución de esta categoría en todas las instancias, mostrando el valor medio y la desviación de los datos. Es útil para observar los puntos máximos (en frecuencia) y en donde tienden a acumularse mas los datos gracias a las aproximaciones con, generalmente, una gráfica normal.

Rank: Filtra las características mas relevantes de la base de datos para posteriormente ignorar el resto, ésto se define gracias a diferentes valores como la ganancia de información, ANOVA (o análisis de varianza), Gini (un coeficiente que determina la igualdad de la distribución), ReliefF (busca la similitud y diferencia entre categorías para determinar un rango de relevancia), entre otros.

Box Plot: Pinta una figura de “cajas” de acuerdo a una categoría seleccionada; la forma de entender este gráfico es así: La gráfica se divide en 4 quartiles, el 2º (el de en medio) representa como una línea el punto medio de la distribución de los datos en la categoría seleccionada; a ambos lados están los quartiles 1 y 3, que en forma de rectángulo muestran lo equivalente al 50% de los datos, los que siguen en la distribución antes y después de la media; por último, los cuartiles 0 y 4, pintados como una línea, representan los datos con menor relevancia en la distribución.

1. **Pca\_example**

Nuevamente, para evitar repeticiones, en este caso vamos a tener 3 widgets, obviando Data Table, Data Table(1) y Scatter Plot:

Datasets: Repetimos la base de datos en línea de las características de un conjunto de vinos.

Paint Data: Nos permite crear una base de datos de 2 variables solo con pintar puntos en un plano de 2 dimensiones.

PCA: Este widget funciona para analizar la variabilidad de 2 características en una base de datos, permite ver la proporción de varianza en los datos para reconocer patrones y relaciones entre las 2 características tomadas con la creación de un gráfico con la menor cantidad de componentes posibles. Normalmente sirve para reducir la dimensionalidad en una base de datos de gran tamaño. La gráfica genera 2 líneas, la roja inferior indica la variabilidad de los componentes individuales, y la verde superior la variabilidad acumulativa entre ambos componentes

**Bibliografía:**

* Orange Visual Programming, Orange Data Mining 2015, [Enlace](https://orange3.readthedocs.io/projects/orange-visual-programming/en/master/index.html)