Taller Certificacion D&A

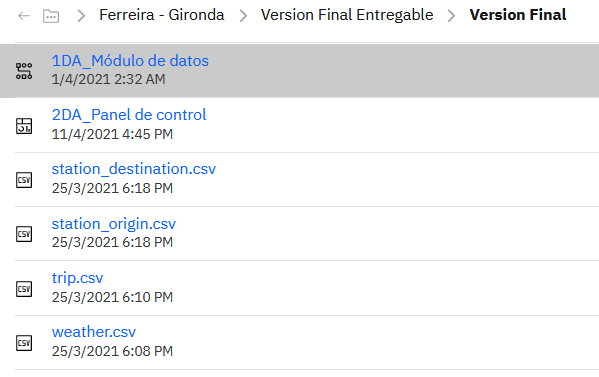
Resolución

El presente documento refiere a la descripción en detalle de los criterios, análisis y desarrollo de los ejercicios propuestos para obtener la certificacion en Introduccion a Data & Analytics :

# Sobre los archivos en COGNOS

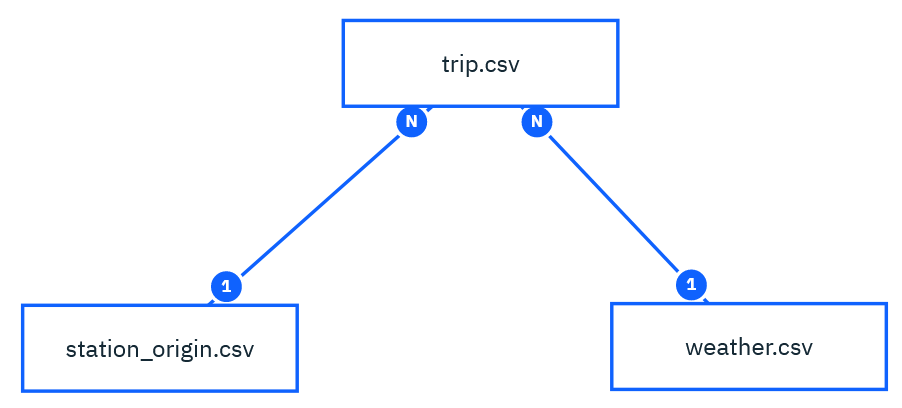
En la herramienta Cognos Analytics, dentro de la sección “Contenido del equipo” el directorio “Ferreira- Gironda > Version Final Entregable > Version Final” contiene los objetos del desarrollo del “modulo de datos” , “panel de control” y “origenes de datos”:

Link - <https://quanam-da-training-ca.quanam.com:9300/bi/?perspective=home&folder=.public_folders%2FTaller%2BFinal%2FFerreira%2B-%2BGironda%2FVersion%2BFinal%2BEntregable>

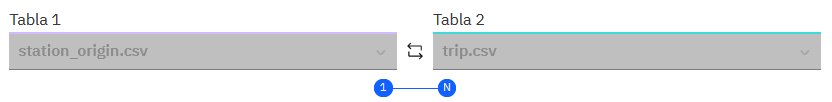


## 1da Módulo de datos

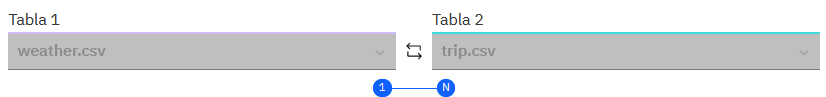
Refiere a el modelo de datos y las relaciones establecidad entre los origenes de datos de los tres diferentes archivos csv.



1ra relación



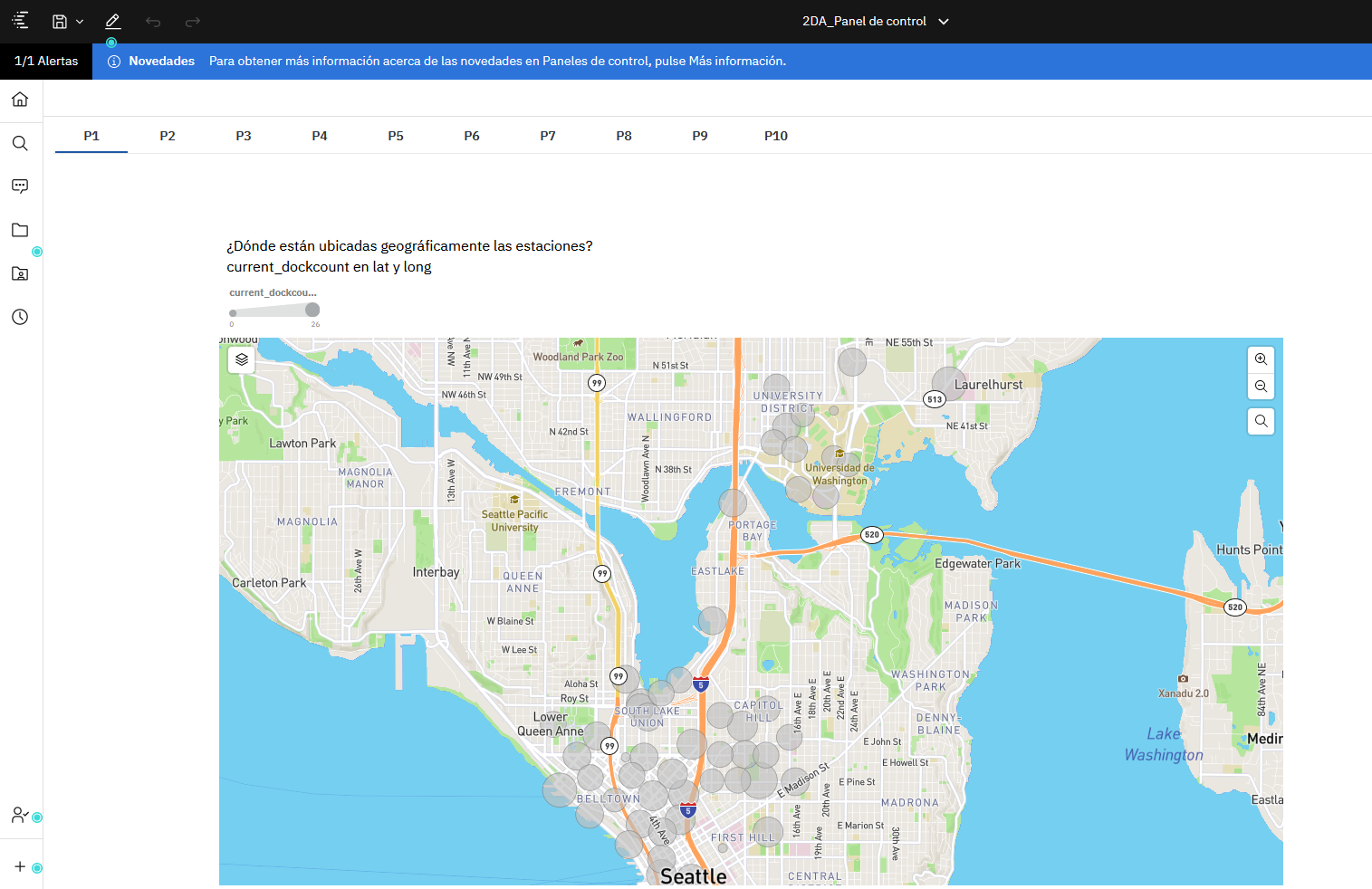
2da relación



Para la 3ra relación también se toma en cuenta que “station\_destination” tiene los mismos registros que “station\_origin” por lo que se obvia esta relación.

# Business Intelligence:

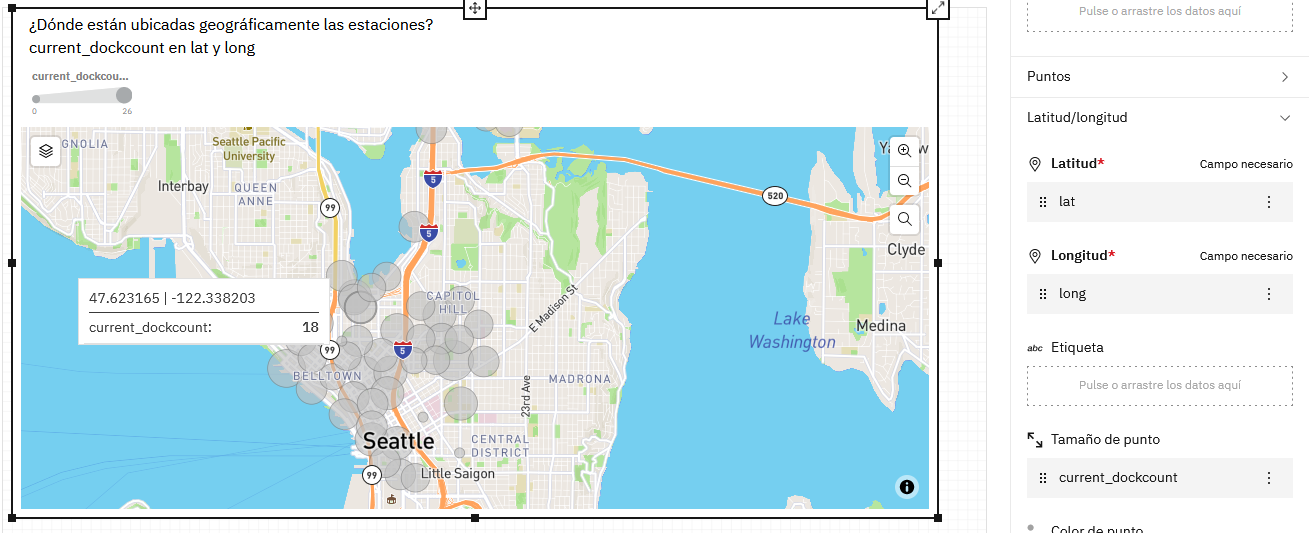
Para el desarrollo de la primera parte referente a BI se desarrolló un Módulo de datos que contiene las respuestas a las cuestiones presentadas en el Taller , separadas por fecha:



Comencemos por:

## 1 ¿Dónde están ubicadas geográficamente las estaciones?

El panel de control “P1” refiere a la referenciación de puntos de latitud y longitud propias de la tabla “estations.csv” . Para hacerlo mas dinámico se presenta en formato de grafico “calles” y se verifica que todos los puntos enmarcan a Seatlle



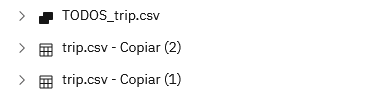
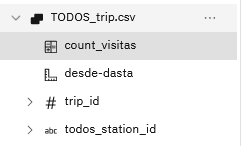
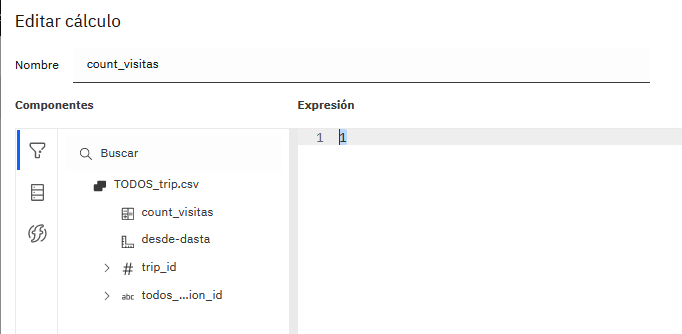
## 2 ¿Cuáles son las 10 estaciones más utilizadas? ¿y las menos utilizadas?

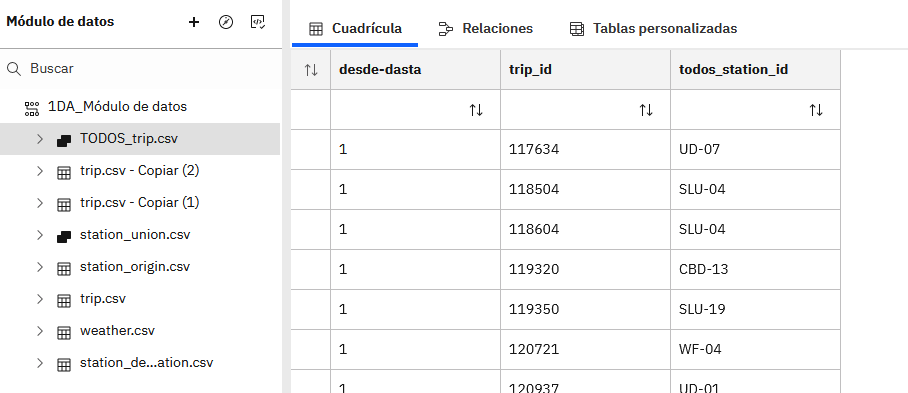
El panel de control “P2” refiere al panel de control:



El panel de control “P2” refiere a las 10 estaciones mas utilizadas. La idea principal es tener una tabla auxiliar que contemple los atributos “to\_station\_id” y “from\_station\_id” de la tabla “trip.csv” como **diferentes** ocurrencias. Esto debido a que se considera como una ocurrencia todos los “to\_station\_id” y una segunda ocurrencia el “from\_station\_id”

Para este fin, se armo una tabla auxiliar “TODOS\_trip.csv” resultado de unir dos tablas temporales “trip.csv- Copiar(2)”y “trip.csv- Copiar(1)””. Éstas últimas refieren a tablas independientes extraídas de “trip.csv” tomando en cuenta para la primera y segunda tabla los atributos “to\_station\_id” y “from\_station\_id” respectivamente, contemplando un cálculo llamado “count\_visitas” con el valor “1”que servirá para validad el conteo de visitas en la estacion.De esta forma conseguimos contemplar en una misma tabla “TODOS\_trip.csv” las ocurrencias de estaciones utilizadas tanto como **origen** como **destino**.

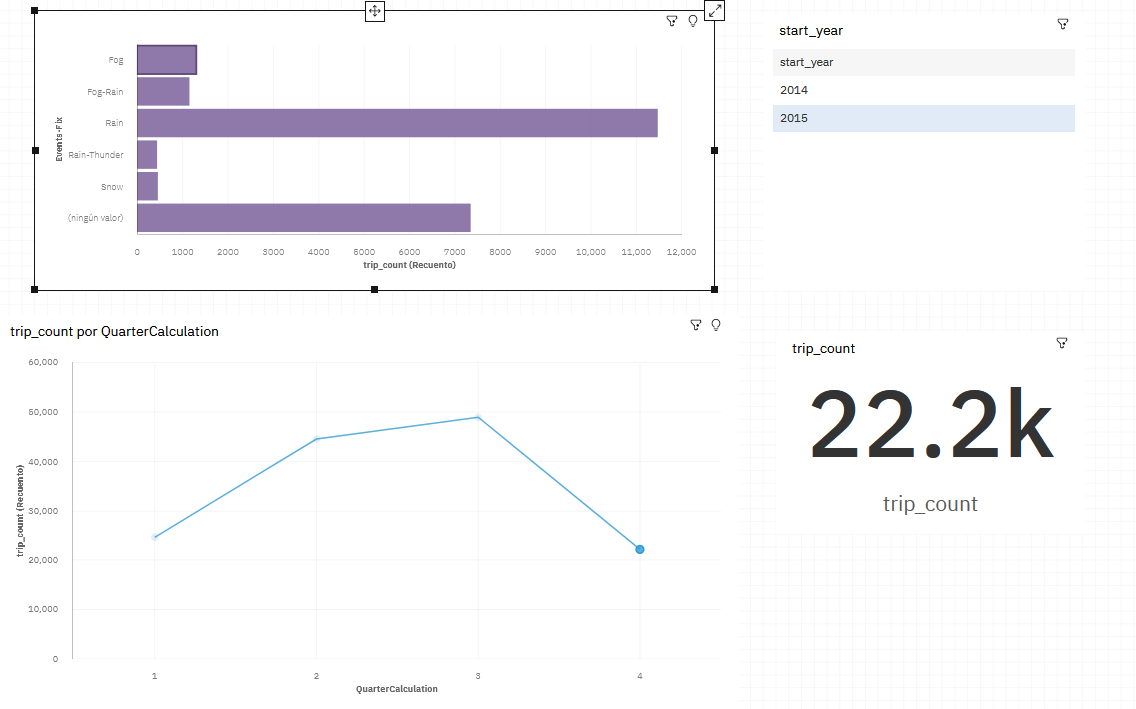
 **->** 



Una vez teniendo esta tabla “TODOS\_trip.csv”, el panel de control propuesto contempla al lado izquierdo los diez station\_id con mayor frecuencia representadas por el tamaño de las burbujas(count). Por el lado derecho muestra los diez station\_id con menor frecuencia (ordenadas descendentemente). Tomar en cuenta que la gráfica permite visualizar la cantidad exacta que veces visitadas al acercarse con el mouse.

## 3 ¿El clima afecta al uso de las bicicletas? ¿De qué manera?

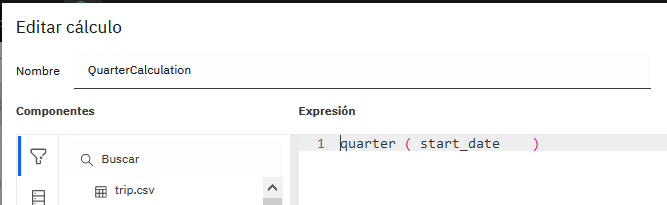
El panel de control “P3” refiere al panel de control:



* El primer grafico hace referencia a un grafico de barras en donde tuvo que crear un “grupo de datos” llamada “Event-Fix” que refería a la columna “Events” con la modificación de que “Fog, Rain” es equivalente a “Fog- Rain”, “Rain, Snow” es equivalene a “Rain- Snow” entre otros similares.

|  |  |
| --- | --- |
| Barras | Event-Fix |
| Longitud | Trip\_count |

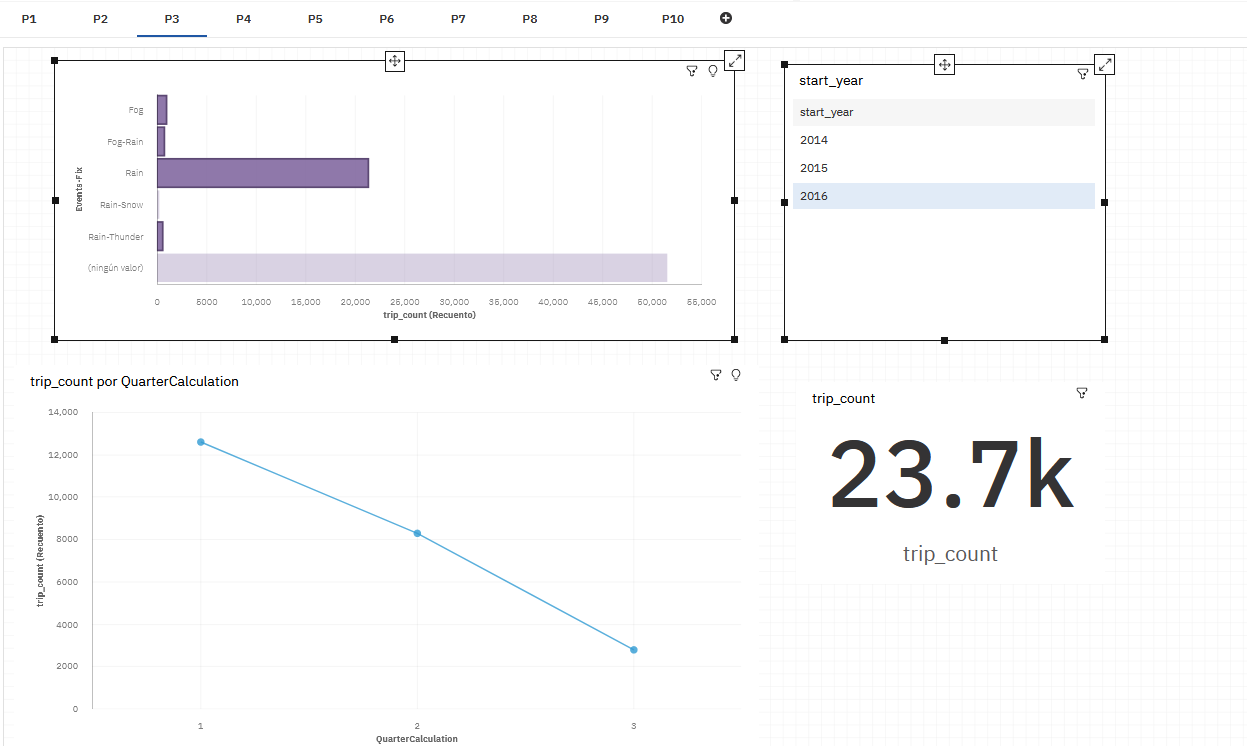
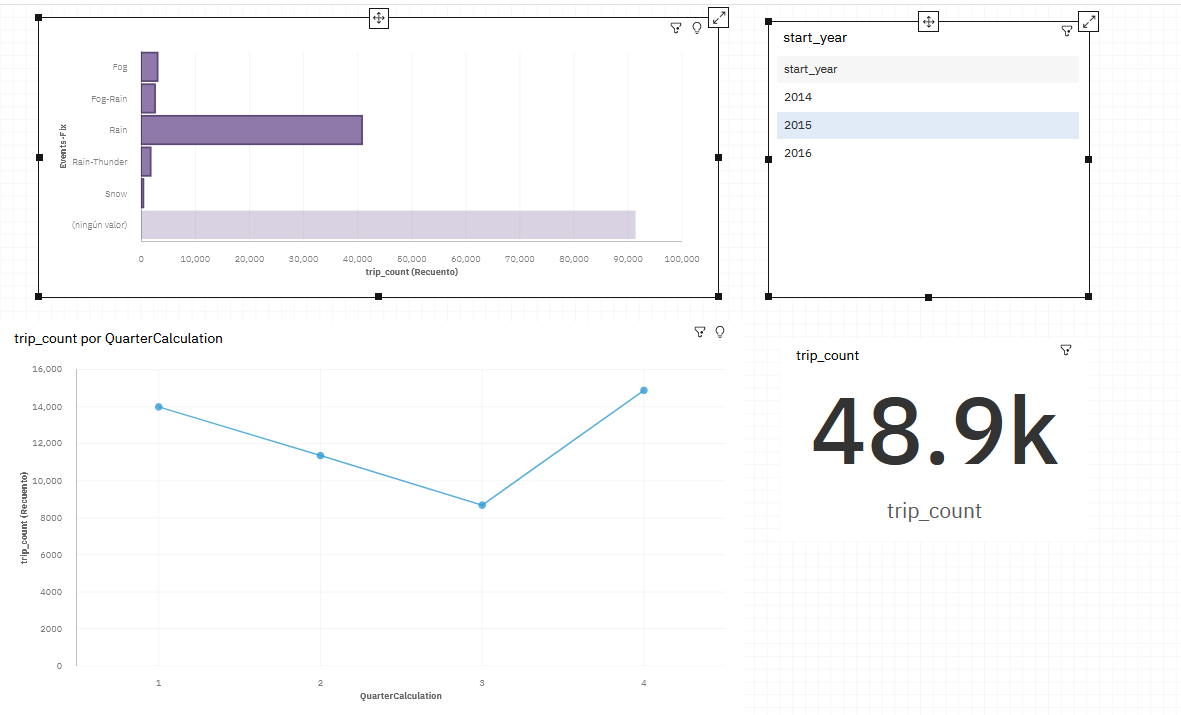
* El segundo refiere a un selector de años para contemplar la variación durante gestiones.
* El tercero refiere a un grafica de línea que muestra la cantidad de viajes contemplados por trimestres de año(quarter). Para este fin se creóun calculo ”QuarterCalculation” que refiere a



|  |  |
| --- | --- |
| Eje x | QuarterCalculation |
| Eje y | Trip\_count |

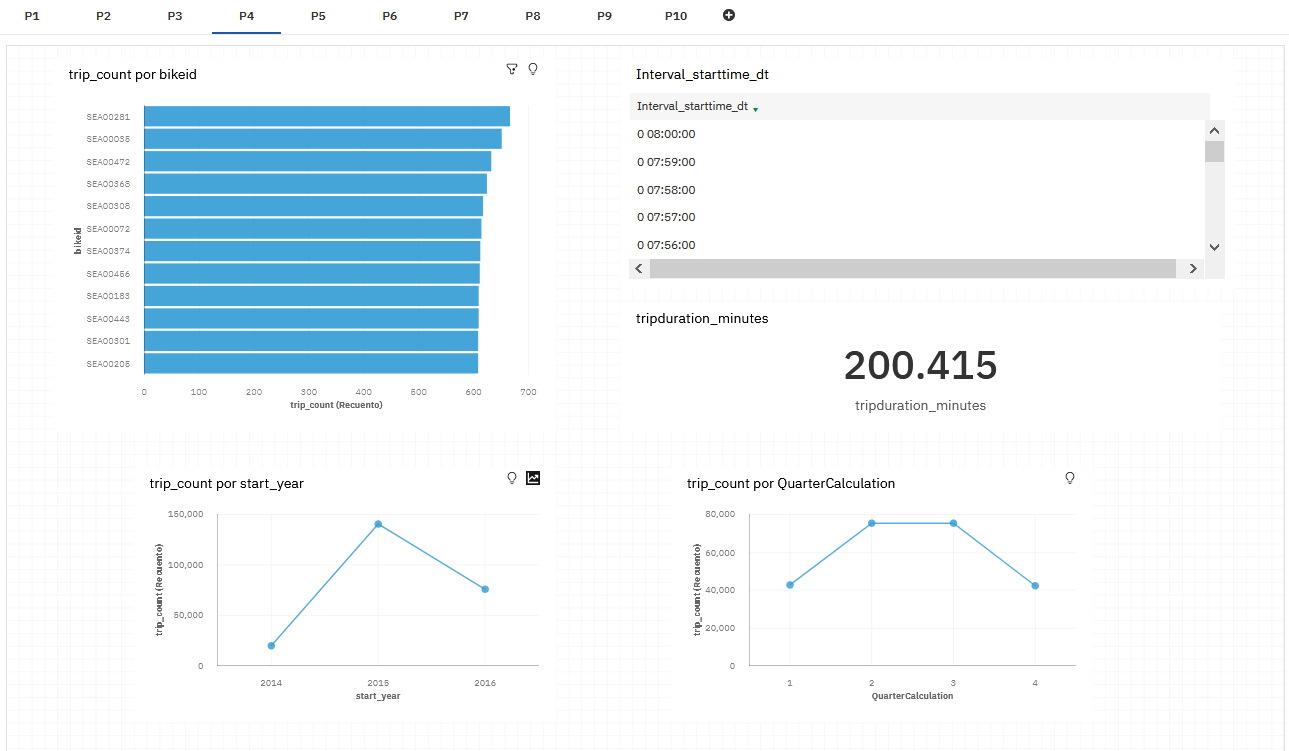
* El cuarto refiere al total de viajes

***Respuesta*** Los gráficos muestran como varia al uso de bicicletas (nro de viajes totales, numero de viajes por trimestre, por año ) contemplando el clima. De aquí pudimos explorar que: Dado que el valor “ningún valor” para el atributo de “Events” representa un día “Normal”, existe mayor cantidad de viajes en todos los años y que el tercer trimestre (julio-septiembre) tiene mayor demanda. Por el contrario, cuanto existe un evento fuera de lo normal, se ve que hay mayor afluencia de utilizar las bicicletas en días lluviosos entre diciembre y enero. Considerando los resultados de las últimas gestiones vemos que las cantidades de viajes por año va disminuyendo, lo que significaría que los días son menos normales en los últimos años, esto se evidencia por lo menos en los dos primeros trimestres en las gráficas.



## 4 ¿Se están utilizando las bicicletas? ¿Cuál es la duración promedio de los viajes?

El panel de control “P4” refiere al panel de control:



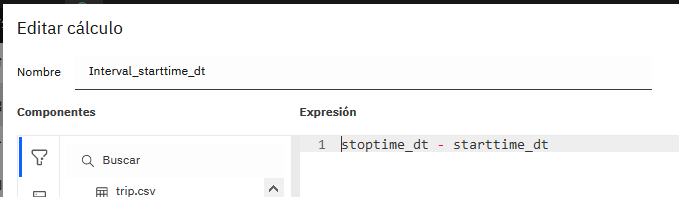
* **trip\_count por bikeid**

El primer grafico hace referencia a un gráfico de barras que refiere a conteo de los viajes por bicicletas ordenados descendentemente:

|  |  |
| --- | --- |
| Barras | Bike\_id |
| Longitud | Trip\_count |

* **Interval\_starttime\_dt**

El segundo refiere al intervalo de tiempo o duración en minutos del viaje de aquellas bicicletas ordenadas de forma descendente. Para este fin, además de explorar que el atributo “tripduration” refería a datos erróneos, se tomó en cuenta un nuevo cálculo “Interval\_starttime\_dt”



* **tripduration\_minutes**

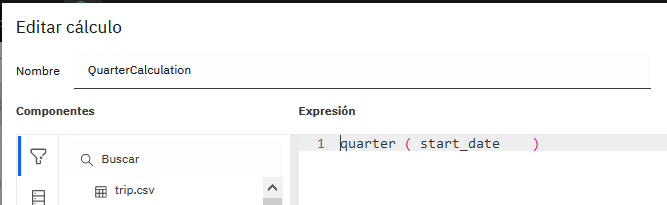
Muestra la duración en minutos total

* **trip\_count por start\_year**

El tercero refiere a un grafica de línea que muestra la cantidad de viajes contemplados por año.

* **trip\_count por QuarterCalculation**

El cuarto refiere a un grafica de línea que muestra la cantidad de viajes contemplados contemplados por trimestres de año(quarter). Para este fin se creóun calculo ”QuarterCalculation” que refiere a:

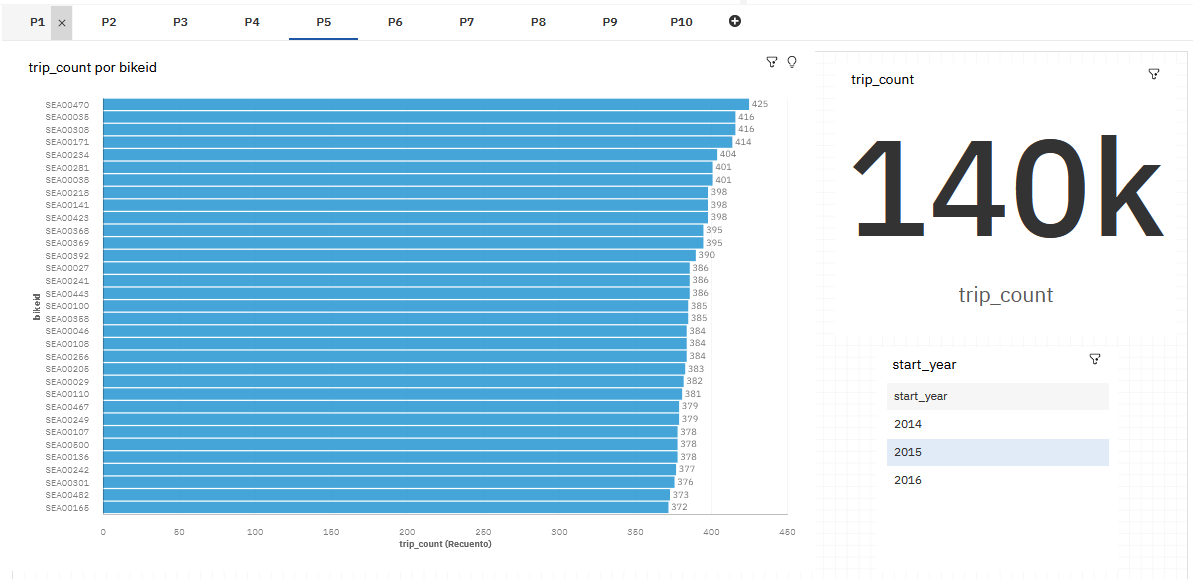


|  |  |
| --- | --- |
| Eje x | QuarterCalculation |
| Eje y | Trip\_count |

***Respuesta*** Si se están usando las bicicletas, el panel de control permite interactura en el tiempo cuantas de ellas se estuvieran utilizando , ya sea por anhio , por trimestre y/o el top de las mas utilizadas mostrando la duración de tiempo. La duración promedio de los viajes refiere al tiempo de duración de los viajes

## 5 ¿Cuál fue la bici más utilizada?

El panel de control “P5” refiere al panel de control.



* **trip\_count por bikeid**

El primer grafico hace referencia a un gráfico de barras que refiere a conteo de los viajes por bicicletas ordenados descendentemente:

|  |  |
| --- | --- |
| Barras | Bike\_id |
| Longitud | Trip\_count |

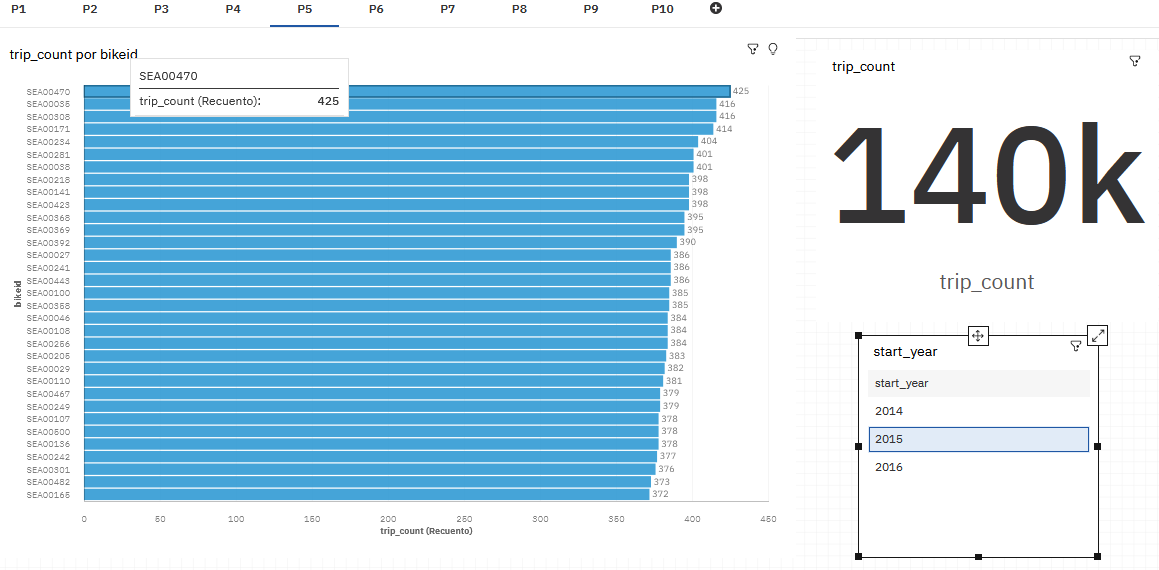
* **trip\_count**

El segundo refiere al conteo de viajes por las bicicletas seleccionadas

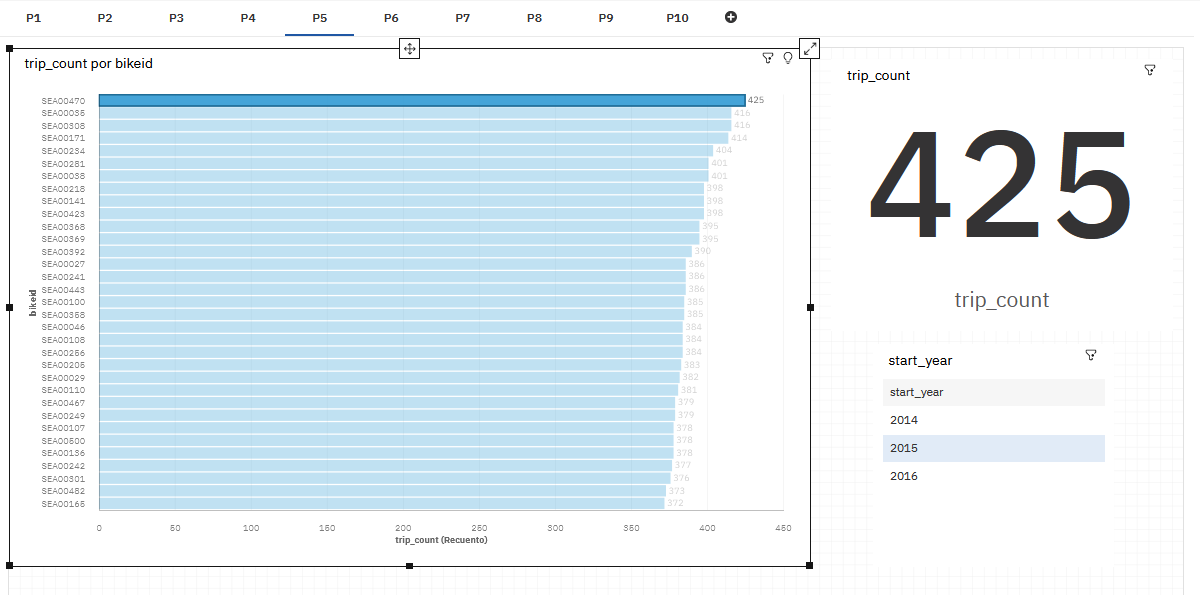
* **startyear**

Muestra un selector de años

***Respuesta*** El panel de control permite visualizar cuales son las bicicletas mas utilizadas teniendo un selector (si se prefiere) por años. Es así si que si seleccionamos de la gestión 2015, tendremos un total de 425 viajes por la bicicleta SEA00470 ademas que entre todas los viajes por las bicis durante la gestión 2015 hacen un total aproximado de 140000 viajes

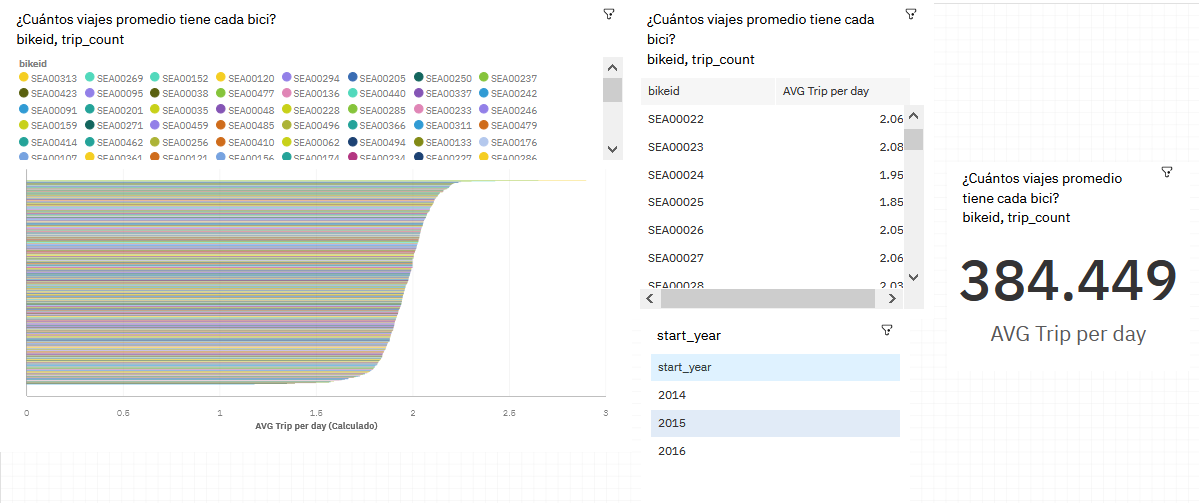


Si seleccionamos solo la bici en cuestión:



## 6 ¿Cuántos viajes promedio tiene cada bici?

El panel de control “P6” refiere al panel de control.



* **bikeid, trip\_count - barras**

El primer grafico hace referencia a un gráfico de barras que refiere a los viajes promedio por cada bicicleta utilizando el cálculo de nombre ”AVG Trip per day”:

|  |  |
| --- | --- |
| Barras | Bike\_id |
| Longitud | Trip\_count |



* **bikeid, trip\_count- tabla lista**

El segundo refiere los mismos datos que el anterior en formato tabla.

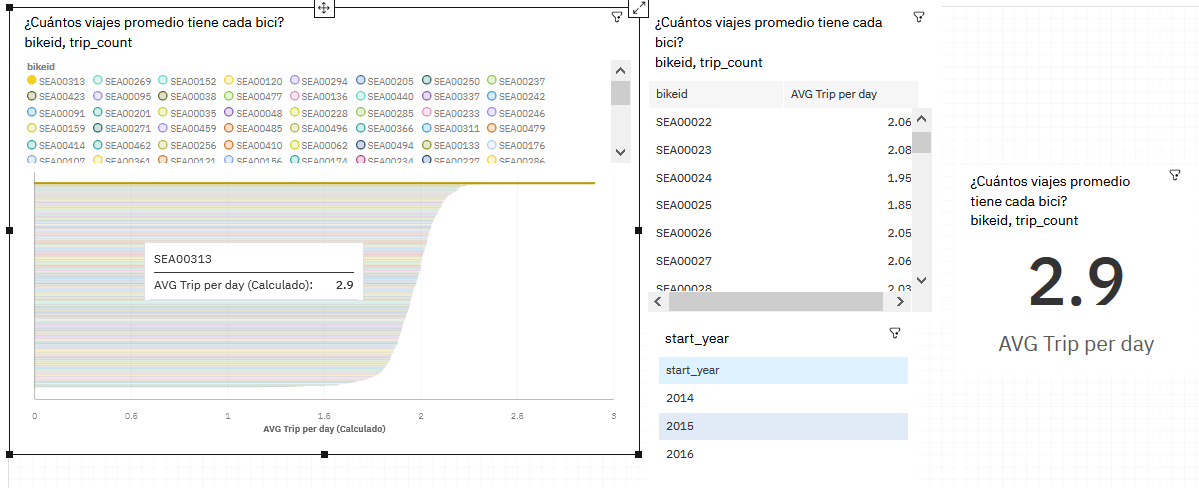
* **startyear**

Muestra un selector de años

* **bikeid, trip\_count Total**

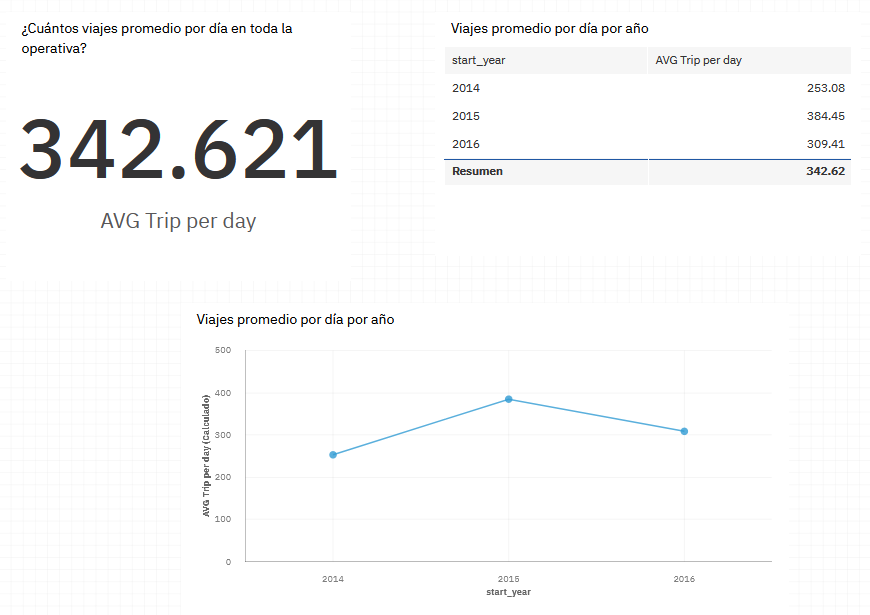
Muestra un resumen de los viajes promedio por día totales.

***Respuesta*** El panel de control permite visualizar cuantos viajes promedio realiza cada bicibleta gracias al cálculo previo realizado “AVG Trip per day” por años seleccionado. Así si seleccionamos el 2015 el promedio de viajes por día es 2.9 para la bicicleta SEA00313, en el panel de control



## 7 ¿Cuántos viajes promedio por día?

El panel de control “P7” refiere al panel de control.



* **¿Cuántos viajes promedio por día en toda la operativa?** -**Resumen**

El primer grafico hace referencia a un resumen del cálculo ”AVG Trip per day” sin tomar en cuenta las bicicletas contempladas en la anterior pregunta:



* **Viajes promedio por día por año – TABLA**

Muestra los viajes promedio por día por años en formato tabla

|  |  |
| --- | --- |
| Columnas | Start\_year  AVG Trip per day |

* **Viajes promedio por día por año - lineas**

Muestra los viajes promedio por día por años en formato de lineas

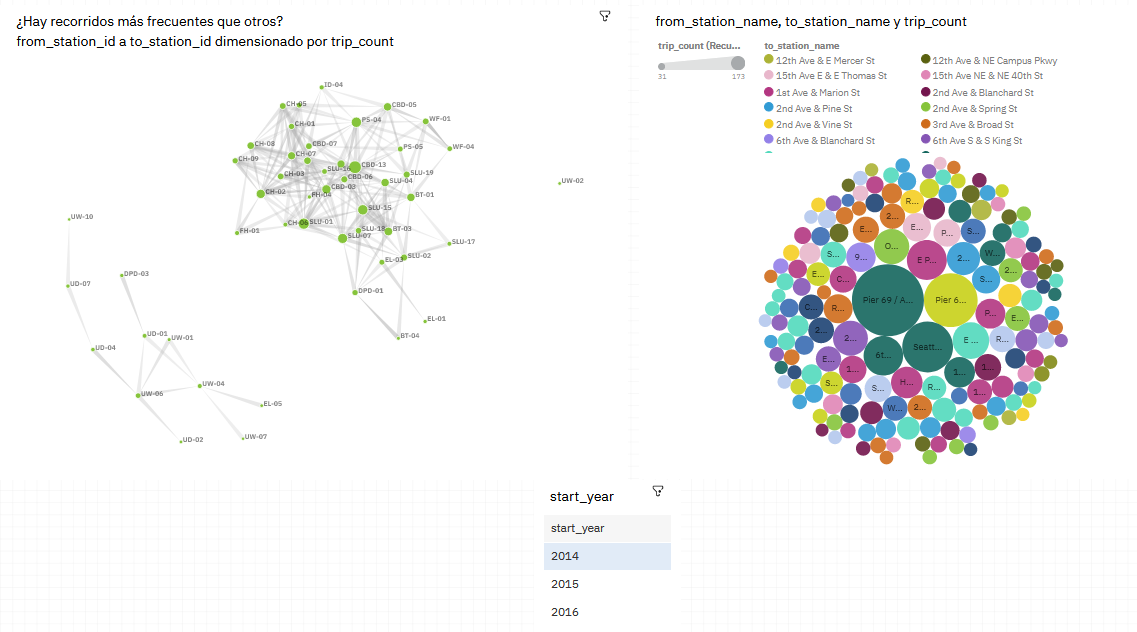
***Respuesta*** El panel de control permite visualizar cuantos viajes promedio realiza cada día gracias al cálculo previo realizado “AVG Trip per day”.

Así si seleccionamos el 2016 el promedio de viajes por día es 309.41, así mismo visualizamos que la gestión 2015 tuvo mayor afluencia de promedio de viajes por día , en el panel de control:



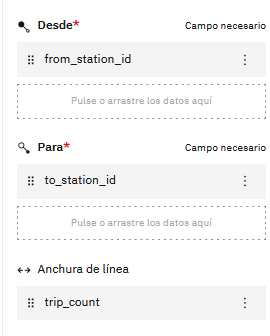
## 8 ¿Hay recorridos más frecuentes que otros?

El panel de control “P8” refiere al panel de control.



* **from\_station\_id a to\_station\_id dimensionado por trip\_count**

El primer grafico hace referencia a un grafico “grafos” que permite visualizar las cantidades de relaciones que existieran entre “from\_station\_id” a “to\_station\_id” haciendo referencia a anchura de la relación al conteo de viajes. Por lo que se muestra con líneas anchas(gruesas) las relaciones de “from\_station\_id” a “to\_station\_id” mas frecuentes.



* **from\_station\_name, to\_station\_name y trip\_count**

El segundo gráfico hace referencia a un grafico “burbujas” contemplando la estación de origen mas utilizada

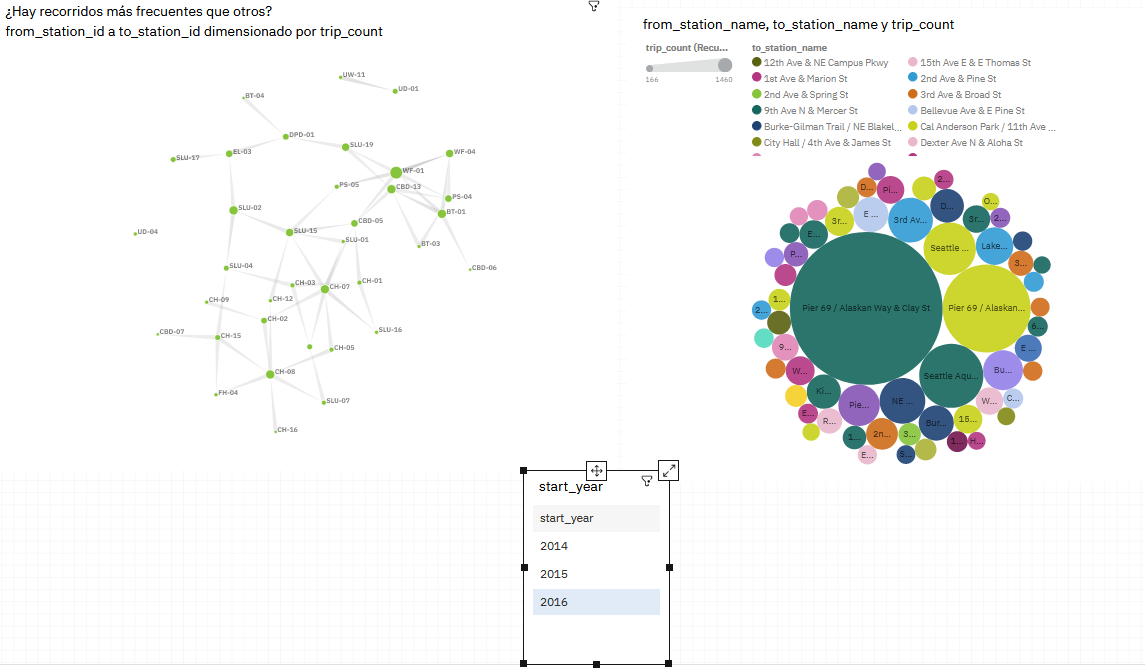
|  |  |
| --- | --- |
| Columnas | Start\_year  AVG Trip per day |

* **start-year**

Muestra selector de años

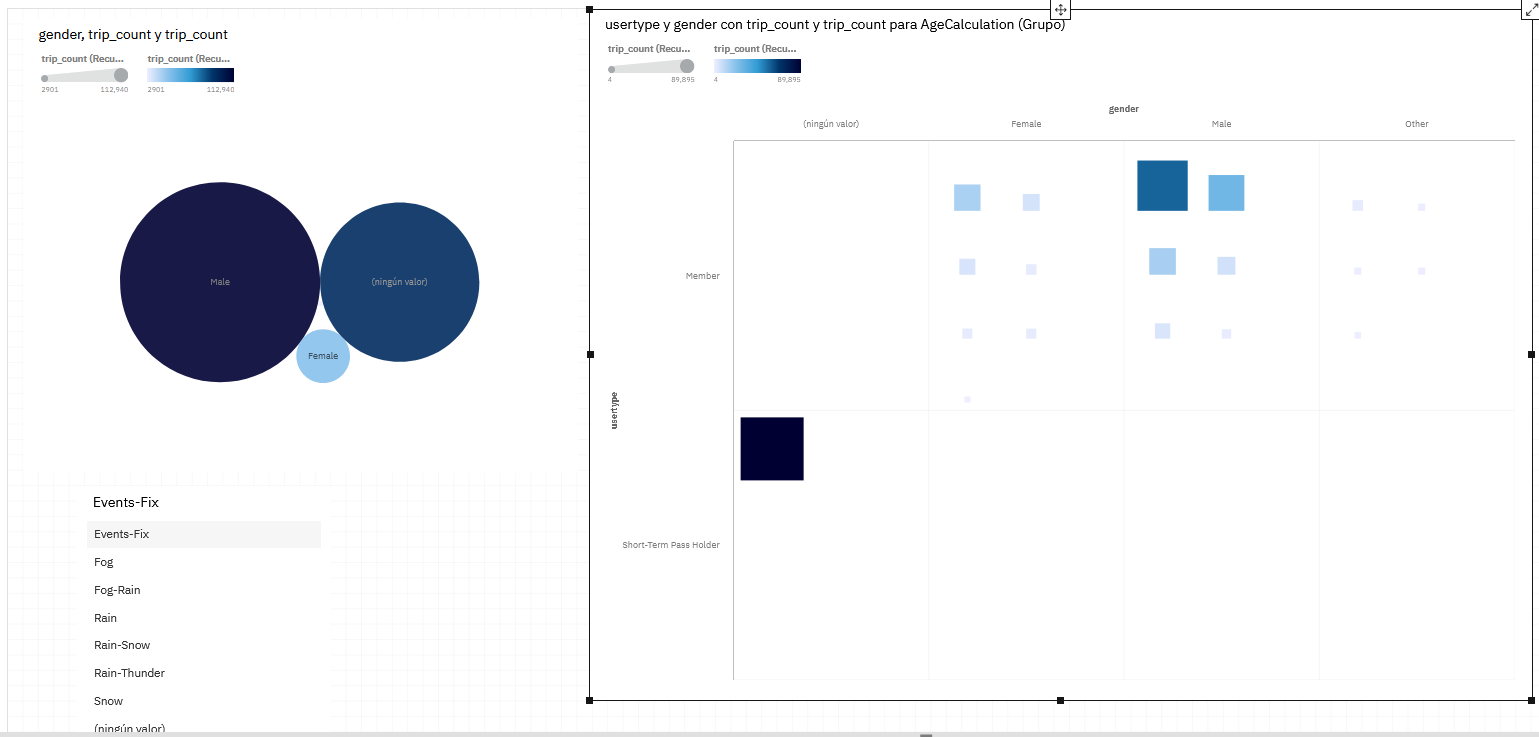
***Respuesta*** Si existe recorridos mas frecuentes que otros. El panel de control propuesto permite visualizar estas frecuencias y si se requiere por años.

El anio 2016:



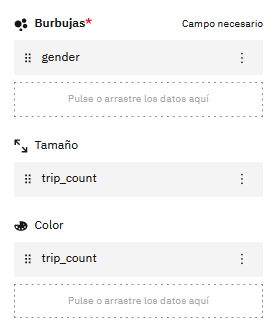
## 9 ¿Podemos conocer mejor a nuestros usuarios? ¿Qué público es el que nos utiliza?

El panel de control “P9” refiere al panel de control.



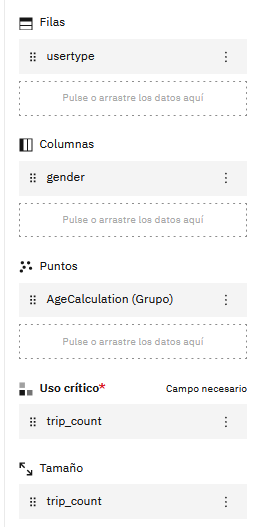
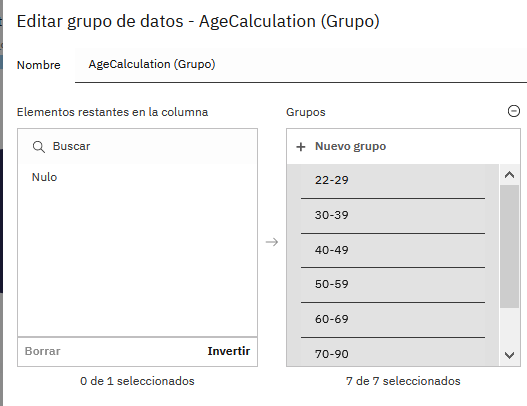
* **gender, trip\_count y trip\_count**

El primer grafico hace referencia a “burbujas” que permite visualizar la cantidad de viajes con relación al género (con color y tamaño para referenciar aquellas con mayor ocurrencia) .



* **usertype y gender con trip\_count y trip\_count para AgeCalculation (Grupo)- MAPA DE USO CRITICO**

El segundo gráfico hace referencia a un grafico “mapa de uso critico” contemplando el “uso critico” como al grupo de datos construido previamente “AgeCalculation”.

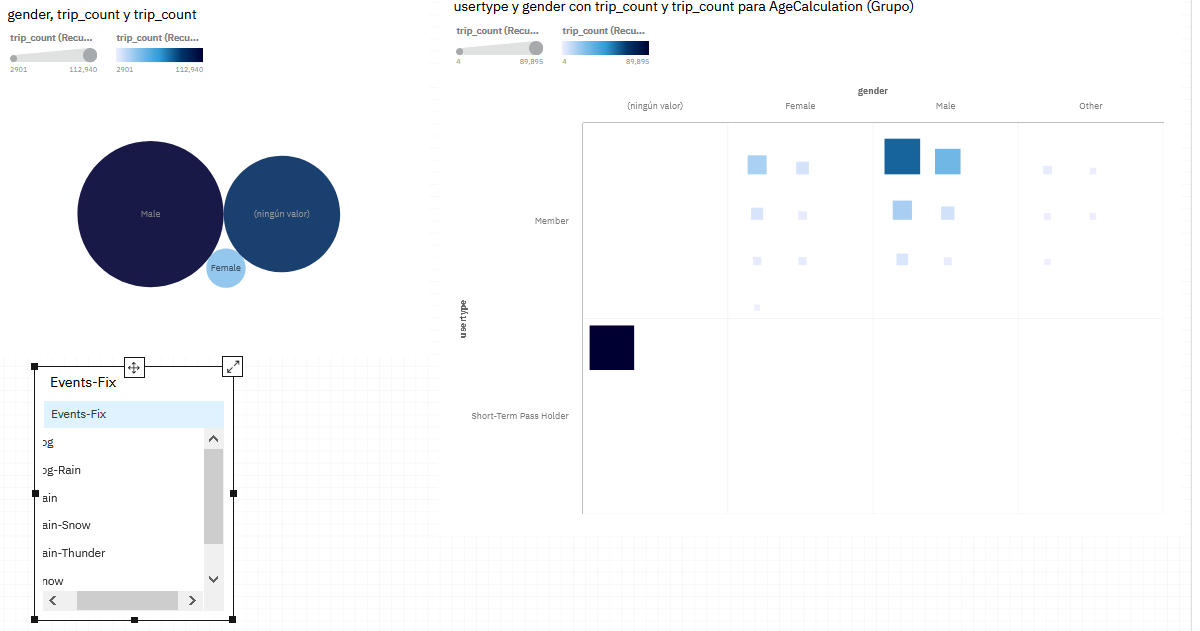
|  |  |
| --- | --- |
| Filas | usertype |
| Columnas | gender |
| Puntos | AgeCalculation |
| Uso Critico | Trip\_count |
| Tamaño | Trip\_count |

* **Events\_Fix**

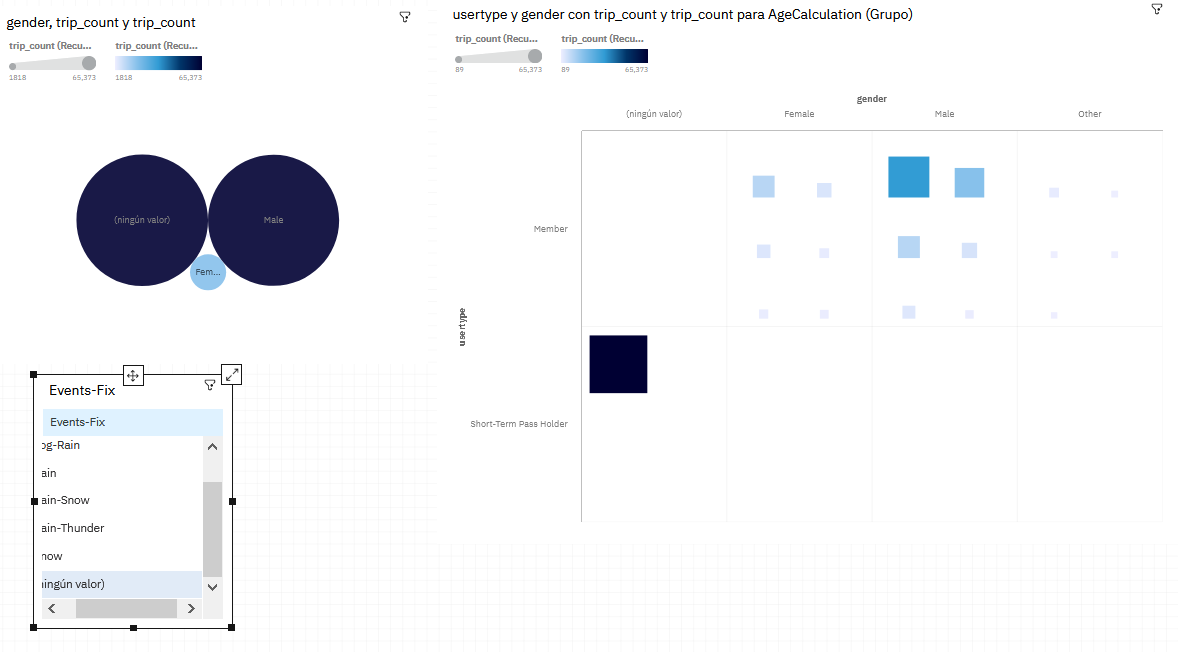
Muestra selector de eventos

***Respuesta*** Si podemos conocer mejor a los usuarios, por ejemplo :

* *Con referencia al primer gráfico de burbujas podemos decir que los usuarios con mayor ocurrencia de realizar viajes son los de genero masculino, son contar con las personas que no quisieron dar información acerca de ello.*



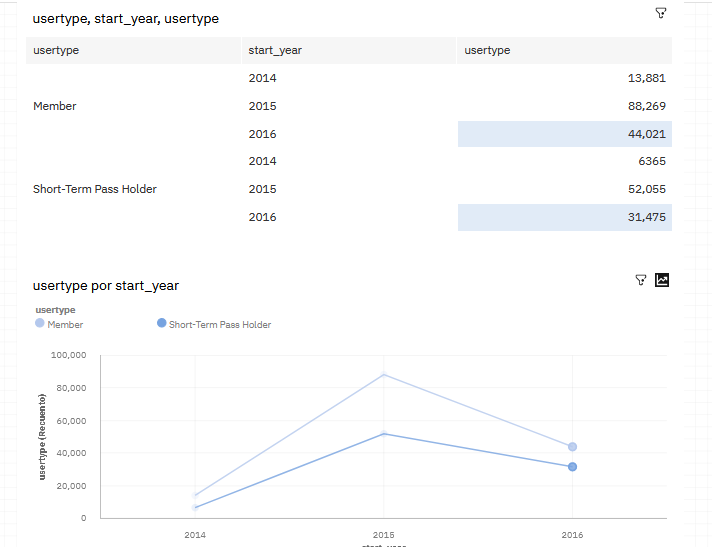
* *También es posible visualizar como son los usuarios en días normlaes o días con eventos como lluviosos. En días normales existe mayor afluencia de viajes por usuarios del tipo “ShortTerm Pas Holder”y son los mayormente no dan datos de su género*



* *También es posible visualizar la cantidad de viajes realizadas por rango de edad contemplando a los cuadrados de mayor tamaño aquellos que tienen mayor cantidad de viajes además del color (aquellos más oscuros son aquellos que mayor cantidad de viajes realizaron ).*

## 10 cómo evolucionó a través de los años la cantidad de usuarios registrados

El panel de control “P10” refiere al panel de control.



* La tabla de arriba muestra una tabla relacionada entre las cantidades de viajes por “userType” en los años. La grafica de abajo refiere a los mismos datos.

**Respuesta** La cantidad de usuarios registrados evolucionó deacuerdo a los a;os según el detalle: La cantidad de viajes de miembros como no miembros subieron relativamente en las ultimas gestiones , comparando con las gestiones pasadas, la gestión 2015 tuvo mayor afluencia de uso de bicicletas, sin contar que el 2016 sobre exoste datos de los nueve primeros meses.

# DataScience - Limpieza y exploración:

Para el desarrollo de la primera parte referente a DataScience se aplicó, desarrolló y ejecutó código python pyspark en IBM Watson.

Cabe destacar que se adjuntan los notebook en formato “ipynb” y “html”, de esta manera no es necesario utilizar ninguna aplicación particular para su revisión, más que un navegador.

Se adjunta a continuación de los dataframes generados

**Schema Join Trips-Weather**

|-- Date: date (nullable = true)

|-- trip\_id: integer (nullable = true)

|-- starttime: timestamp (nullable = true)

|-- stoptime: timestamp (nullable = true)

|-- bikeid: string (nullable = true)

|-- from\_station\_name: string (nullable = true)

|-- to\_station\_name: string (nullable = true)

|-- from\_station\_id: string (nullable = true)

|-- to\_station\_id: string (nullable = true)

|-- usertype: string (nullable = true)

|-- gender: string (nullable = true)

|-- birthyear: integer (nullable = true)

|-- age: integer (nullable = true)

|-- quarter\_date: integer (nullable = true)

|-- age\_range: string (nullable = true)

|-- trip\_duration\_seconds: long (nullable = true)

|-- trip\_duration\_minutes: double (nullable = true)

|-- Max\_Temperature\_F: integer (nullable = true)

|-- Mean\_Temperature\_F: double (nullable = true)

|-- Min\_TemperatureF: integer (nullable = true)

|-- Max\_Dew\_Point\_F: integer (nullable = true)

|-- MeanDew\_Point\_F: integer (nullable = true)

|-- Min\_Dewpoint\_F: integer (nullable = true)

|-- Max\_Humidity: integer (nullable = true)

|-- Mean\_Humidity: integer (nullable = true)

|-- Min\_Humidity: integer (nullable = true)

|-- Max\_Sea\_Level\_Pressure\_In: double (nullable = true)

|-- Mean\_Sea\_Level\_Pressure\_In: double (nullable = true)

|-- Min\_Sea\_Level\_Pressure\_In: double (nullable = true)

|-- Max\_Visibility\_Miles: integer (nullable = true)

|-- Mean\_Visibility\_Miles: integer (nullable = true)

|-- Min\_Visibility\_Miles: integer (nullable = true)

|-- Max\_Wind\_Speed\_MPH: integer (nullable = true)

|-- Mean\_Wind\_Speed\_MPH: integer (nullable = true)

|-- Max\_Gust\_Speed\_MPH: integer (nullable = true)

|-- Precipitation\_In: double (nullable = true)

|-- Events: string (nullable = true)

|-- temporada: string (nullable = true)

**Schema Join Trips-ToStation**

|-- trip\_id: integer (nullable = true)

|-- starttime: timestamp (nullable = true)

|-- stoptime: timestamp (nullable = true)

|-- bikeid: string (nullable = true)

|-- from\_station\_name: string (nullable = true)

|-- to\_station\_name: string (nullable = true)

|-- from\_station\_id: string (nullable = true)

|-- to\_station\_id: string (nullable = true)

|-- usertype: string (nullable = true)

|-- gender: string (nullable = true)

|-- birthyear: integer (nullable = true)

|-- Date: date (nullable = true)

|-- age: integer (nullable = true)

|-- quarter\_date: integer (nullable = true)

|-- age\_range: string (nullable = true)

|-- trip\_duration\_seconds: long (nullable = true)

|-- trip\_duration\_minutes: double (nullable = true)

|-- station\_id: string (nullable = true)

|-- name: string (nullable = true)

|-- lat: double (nullable = true)

|-- long: double (nullable = true)

|-- install\_date: date (nullable = true)

|-- install\_dockcount: integer (nullable = true)

|-- modification\_date: date (nullable = true)

|-- current\_dockcount: integer (nullable = true)

|-- decommission\_date: date (nullable = true)

|-- quarter\_install\_date: integer (nullable = true)

**Schema Join Trips-FromStation**

|-- trip\_id: integer (nullable = true)

|-- starttime: timestamp (nullable = true)

|-- stoptime: timestamp (nullable = true)

|-- bikeid: string (nullable = true)

|-- from\_station\_name: string (nullable = true)

|-- to\_station\_name: string (nullable = true)

|-- from\_station\_id: string (nullable = true)

|-- to\_station\_id: string (nullable = true)

|-- usertype: string (nullable = true)

|-- gender: string (nullable = true)

|-- birthyear: integer (nullable = true)

|-- Date: date (nullable = true)

|-- age: integer (nullable = true)

|-- quarter\_date: integer (nullable = true)

|-- age\_range: string (nullable = true)

|-- trip\_duration\_seconds: long (nullable = true)

|-- trip\_duration\_minutes: double (nullable = true)

|-- station\_id: string (nullable = true)

|-- name: string (nullable = true)

|-- lat: double (nullable = true)

|-- long: double (nullable = true)

|-- install\_date: date (nullable = true)

|-- install\_dockcount: integer (nullable = true)

|-- modification\_date: date (nullable = true)

|-- current\_dockcount: integer (nullable = true)

|-- decommission\_date: date (nullable = true)

|-- quarter\_install\_date: integer (nullable = true)

**Schema Trips**

|-- trip\_id: integer (nullable = true)

|-- starttime: timestamp (nullable = true)

|-- stoptime: timestamp (nullable = true)

|-- bikeid: string (nullable = true)

|-- from\_station\_name: string (nullable = true)

|-- to\_station\_name: string (nullable = true)

|-- from\_station\_id: string (nullable = true)

|-- to\_station\_id: string (nullable = true)

|-- usertype: string (nullable = true)

|-- gender: string (nullable = true)

|-- birthyear: integer (nullable = true)

|-- Date: date (nullable = true)

|-- age: integer (nullable = true)

|-- quarter\_date: integer (nullable = true)

|-- age\_range: string (nullable = true)

|-- trip\_duration\_seconds: long (nullable = true)

|-- trip\_duration\_minutes: double (nullable = true)

**Schema Weather**

|-- Date: date (nullable = true)

|-- Max\_Temperature\_F: integer (nullable = true)

|-- Mean\_Temperature\_F: double (nullable = true)

|-- Min\_TemperatureF: integer (nullable = true)

|-- Max\_Dew\_Point\_F: integer (nullable = true)

|-- MeanDew\_Point\_F: integer (nullable = true)

|-- Min\_Dewpoint\_F: integer (nullable = true)

|-- Max\_Humidity: integer (nullable = true)

|-- Mean\_Humidity: integer (nullable = true)

|-- Min\_Humidity: integer (nullable = true)

|-- Max\_Sea\_Level\_Pressure\_In: double (nullable = true)

|-- Mean\_Sea\_Level\_Pressure\_In: double (nullable = true)

|-- Min\_Sea\_Level\_Pressure\_In: double (nullable = true)

|-- Max\_Visibility\_Miles: integer (nullable = true)

|-- Mean\_Visibility\_Miles: integer (nullable = true)

|-- Min\_Visibility\_Miles: integer (nullable = true)

|-- Max\_Wind\_Speed\_MPH: integer (nullable = true)

|-- Mean\_Wind\_Speed\_MPH: integer (nullable = true)

|-- Max\_Gust\_Speed\_MPH: integer (nullable = true)

|-- Precipitation\_In: double (nullable = true)

|-- Events: string (nullable = true)

|-- temporada: string (nullable = true)

**Schema Stations**

|-- station\_id: string (nullable = true)

|-- name: string (nullable = true)

|-- lat: double (nullable = true)

|-- long: double (nullable = true)

|-- install\_date: date (nullable = true)

|-- install\_dockcount: integer (nullable = true)

|-- modification\_date: date (nullable = true)

|-- current\_dockcount: integer (nullable = true)

|-- decommission\_date: date (nullable = true)

|-- quarter\_install\_date: integer (nullable = true)

## 

## 1 Desafíos iniciales

Es conocer la data, el negocio (escenario y contexto), poder buscar inconsistencias y limpiar registros.

Entre esto podemos destacar los siguientes

* A nivel general
  + Poder asegurar que las fechas estuvieran almacenadas como fechas sin importar el formato que tuvieran (MM/DD/YY, M/DD/YYYY,MM/D/YY,etc)
  + Eliminación de duplicados
  + Formateo de datos general a su correcto formato
* A nivel de Weather
  + Entender porque ciertos registros contaban con nulos, más específicamente
    - Mean\_Temperature\_F
      * En este caso era un registro y se decide calcular, si bien el caso puntual, la temperatura máxima y mínima son las mismas. Se entendió que no era justificable tener un único nulo contando con data para calcular así fuera un resultado obvio
    - Max\_Gust\_Speed\_MPH
      * En este caso se decide permanecer con datos nulos, que posteriormente para almacenar en archivo parquet, se transforma a un tipo así como “SinDato”. Aún así se investigó utilizando fuentes de inumet para entender si era normal que pudieran haber tantos registros con datos faltantes y/o nulos.
        + Para entender más de esto debemos referirnos a las siguientes fuentes

Podemos afirmar que la rafaga de viento es "el aumento breve y repentino de la velocidad del viento respecto a su valor medio, fuente OMM Nº182)" (fuente https://www.inumet.gub.uy/sites/default/files/2018-04/Inumet\_Informe\_postevento\_12abril2018.pdf)

Es un aumento repentino y significativo en las fluctuaciones de la velocidad del viento. La velocidad punta del viento debe alcanzar por lo menos 16 nudos (30 km/h / 18.64 mp/h) y la variación entre los picos y la calma es de por lo menos 10 nudos (18 km/h / 11.18 mp/h). Generalmente la duración es menor de 20 segundos. (fuente <https://www.inumet.gub.uy/institucional/glosario>)

* + Events
    - Encontramos valores repetidos ingresados en formas distintas, se procedió a estandarizar
  + Generación de estación (meteorológicas)
    - Basados en los periodos de estaciones de USA, ya que las estaciones se ubican en Seattle. (fuente <https://www.visittheusa.mx/info/tiempo-y-estaciones-del-ano>)
* A nivel de trips
  + Generación de una columna fecha de tipo Date únicamente, necesario para hacer join contra la data de clima
  + Se buscó registros duplicados y se eliminaron
  + Se buscó registros que tuvieran un stoptime menor al startime, es decir, viajes en lo que se “comienza posteriormente de haber finalizado”. Esto último claramente no era coherente, de tal manera que se eliminan estos registros.
  + Se genera un atributo de tipo string almacenando la fecha como “YYYYMMDD” para posterior manejo alternativo.
* A nivel de gender y birth year
  + Se identifican datos tipo null, el cual es aceptado para el negocio, con previa consulta vía correo electrónico. Entendemos que el cliente simplemente no quiso brindar la información.
* Se genera un atributo categórico para clasificar edad
* Se genera un atributo de cálculo de edad
  + Este se considera en base al cálculo de la diferencia del birth year vs el año en del startime

## 2 Análisis de variables

En cuanto a este apartado destacamos

* Indizar los atributos Events de clima y Género de viajes
* Revisión de distribución de duración en minutos en quintiles, identificando un grupo muy dispar al resto. Dejando abierta la posibilidad a una futura revisión con más buckets para entender más sobre ello.
* Se revisa la progresión anual de temperatura media, temperatura máxima y mínima pudiendo notar que hay una tendencia de aumento en el microclima de Seattle.

## 3 Exploración de estrategias para analizar clima mensualmente

En cuanto a este apartado entendimos crear un KPI propio basado en la cantidad de registros de eventos de clima que tenemos.

En particular nos centramos en que tenemos “días normales” (vacíos originalmente) y el resto son días que cuentan con alguna inclemencia meteorológica. Por lo cual entendemos que podíamos contar con un KPI que fuera cantidad de días normales dividido el total (“días normales” + días con inclemencia meteorológica).

Cabe destacar que se generó un atributo llamado “IsGoodWeather” de tipo integer que tiene dos valores posibles 0 (“cero”) o 1. Se entiende por 1 día normal y 0 por día con inclemencia meteorológica.

Se mostró en una visualización agrupada y ordenada por año y mes.