# Ejercicios del curso de Visualización de Datos

## Integrantes:


➢ David Alberto Rodriguez Muñoz dalberto.rodriguez@udea.edu.co

➢ Victor Alberto Lizcano Portilla alberto.lizcano@udea.edu.co


**Ejercicio 1**

Visita el sitio web https://fathom.info/salaryper/ y analizando la visualización identifica que atributos visuales han sido utilizados en la visualización? ¿A qué variable cuantitativa o categórica se le ha asignado cada atributo visual?

**Respuesta:**

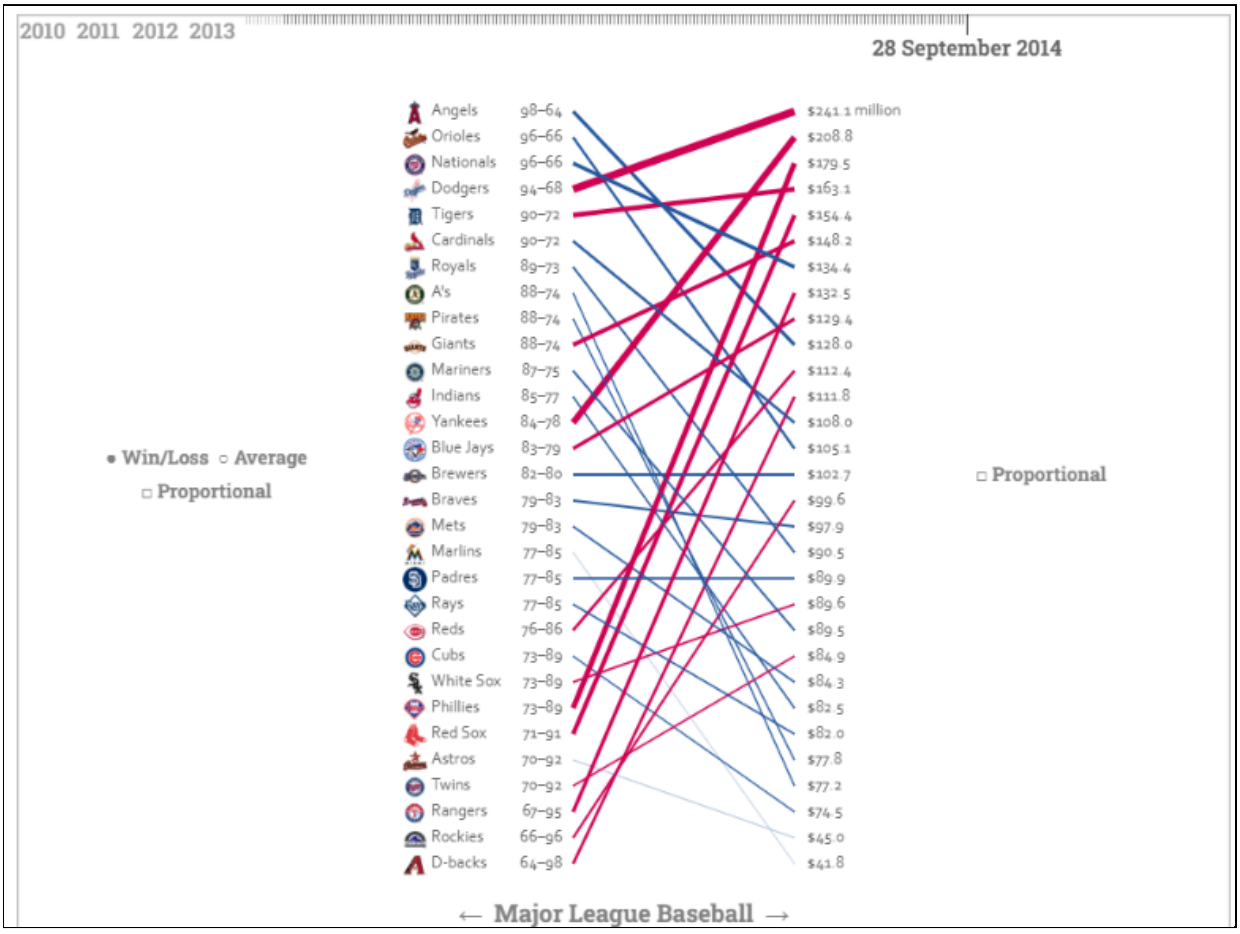


*Figura 1*

Luego de realizar un análisis del gráfico de la figura 1, se tienen los siguientes atributos visuales y la variable a la cual pertenece:

- Color: asignado a la variable categórica que hace alusión al comportamiento positivo o negativo sobre el uso adecuado del dinero.
- Ancho de la línea: Identifica la relación del salario del equipo con los demás, entre más ancho mayor es el salario.
- Orientación: asignado a la variable categoría relacionado con el buen uso del dinero, pendiente positiva, uso inadecuado de los recursos; pendiente negativa, uso favorable de los recursos.
- Posición: utilizada para representar la variable categórica de la liga
- Movimiento: Representa el comportamiento de los equipos en el tiempo.

## Ejercicio 2

Usted trabaja para una aerolínea y es encargado de dirigir a un grupo de desarrolladores de software que están construyendo el nuevo portal web que le permitirá a los usuarios poder comprar boletos y hacer el check-in de sus vuelos.
Su equipo de desarrolladores trabaja en ciclos de 15 días de trabajo. Al primer día del ciclo se hace una reunión en la que usted le presenta todas las tareas que hay por hacer con su respectiva prioridad. Ellos evalúan cada una asignándole una puntuación de cuánto esfuerzo le tomará realizarla, estos puntos están en el rango entre 1 y 100. También estima cuántos días le tomaría hacer la tarea (Tiempo en días).
Las tareas pueden ser de 3 tipos diferentes: desarrollando una *Feature* al portal web se le adiciona una nueva funcionalidad que no existía (por ejemplo, hacer check-in online), desarrollando una *User Story* al portal se le adiciona algo sencillo que lo hace mejor (por ejemplo, que el usuario pudiera imprimir su boleto de vuelo) y solucionando un *Bug* entonces se resuelve un problema que ya existía.
La siguiente tabla muestra los resultados de la reunión para las 9 tareas que usted trajo al grupo ese primer día. Primero tome un tiempo y piense usted como director de grupo que preguntas le interesaría responder con esos datos. Segundo, dibuje la visualización que usted cree que mejor respondería sus preguntas. Recuerde que los datos que muestra la tabla son solo ilustrativos, pueden variar de reunión en reunión.

| *Tipo* | *Prioridad* | *Esfuerzo (puntos)* | *Tiempo (días)* |
|---|---|---|---|
| Feature | Must Have | 30 | 40 |
| Feature | Good | 20 | 40 |
| Feature | Nice to Have | 15 | 20 |
| Bug | Fix ASAP | 2 | 2 |

| | | | |
|---|---|---|---|
| Bug | Fix | 2 | 8 |
| Bug | Fix if Time | 5 | 12 |
| User Story | Must Have | 8 | 10 |
| User Story | Good | 5 | 7 |
| User Story | Nice to Have | 8 | 7 |

**Respuesta**

a. ¿Con qué rapidez o incremento y en qué tiempo se finalizan todas las tareas, si se tienen 1, 2 y 5 desarrolladores?

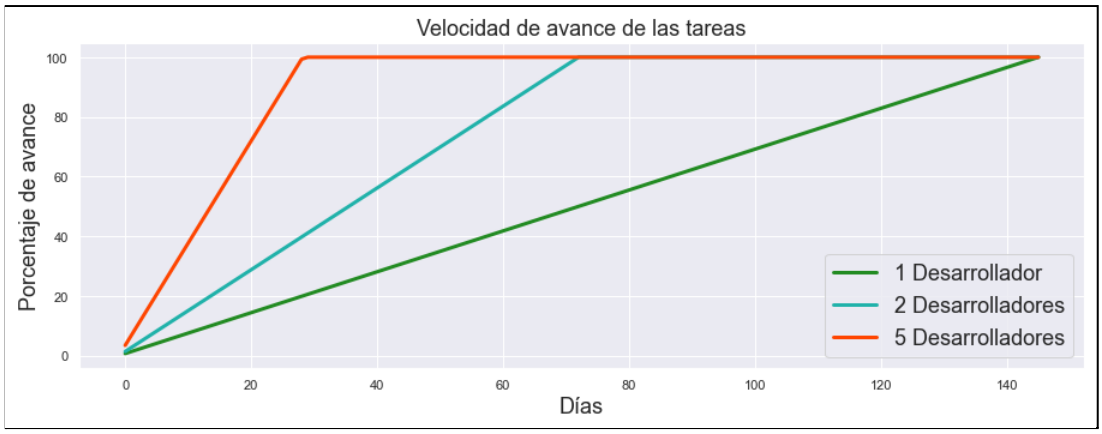


*Figura 2*

Como se puede ver en la figura 2, el gráfico permite identificar la relación entre el número de desarrolladores y el tiempo que tarda en concluir todas las tareas. De forma adicional me permite conocer la rapidez con que se cumplen las tareas dependiendo del número de desarrolladores.

b. ¿Existe una relación entre el tipo de tarea, tiempo en que tarda en desarrollarse y el esfuerzo requerido?

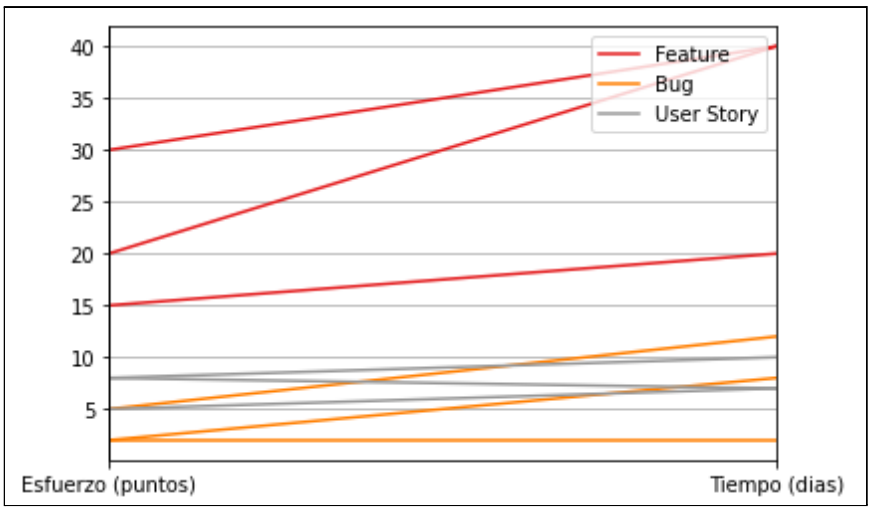


Figura 3

Como se puede ver en la figura 3 existe una correlación positiva entre el esfuerzo que requiere cualquiera de las tres tareas y el tiempo que tarda en desarrollarse, siendo el desarrollo de una nueva característica, la que requiere mayor tiempo y esfuerzo.

c. ¿De las tres posibles tareas desarrolladas, cuál representa el mayor porcentaje de trabajo, cómo se ve reflejado en días y en esfuerzo?

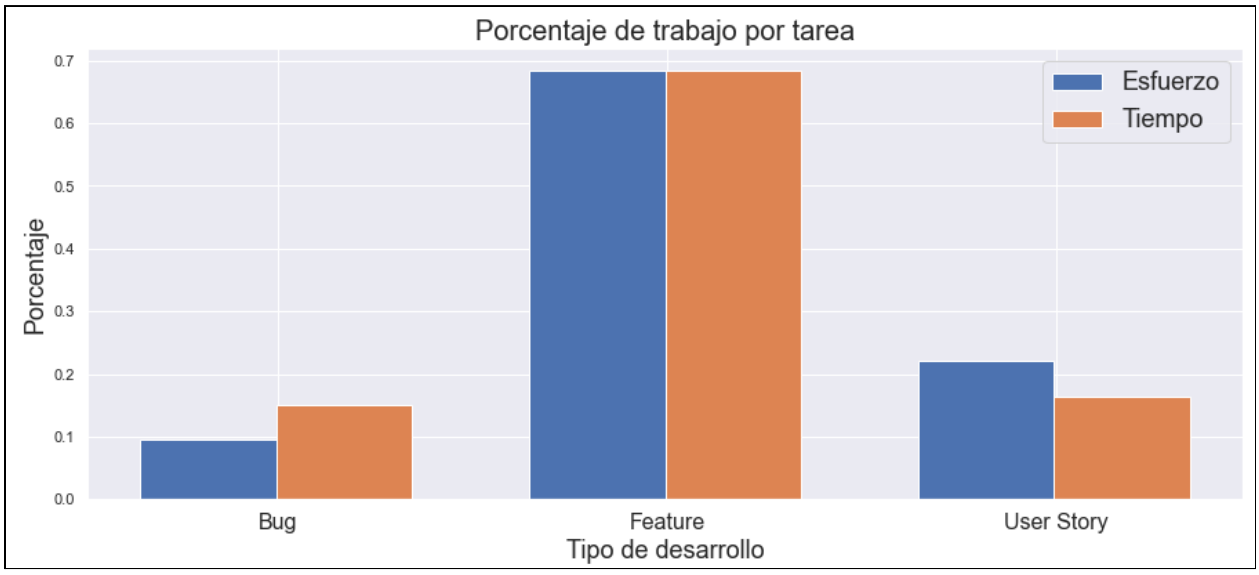


Figura 4

Como se puede ver en la figura 4, la tarea que mayor esfuerzo y tiempo requiere sin duda alguna es el desarrollo de una nueva feature, sin embargo, dar solución a un bug no requiere tanto esfuerzo como tiempo, mientras que atender un user story requiere un mayor esfuerzo en relación con el tiempo.

**Ejercicio 3**

A continuación, se muestran escenarios de la vida práctica para los cuales se necesita que determines cuál es la forma de diseño mas efectiva para comunicar el mensaje cuantitativo. Para llegar a tu respuesta debes ir respondiendo a esta serie de preguntas:

- ¿El mensaje debe ser presentado con un gráfico o una tabla?
- ¿Si es una tabla, podrías hacer un boceto del mismo?
- Si es un gráfico, ¿qué tipo de relación cuantitativa debe ser mostrada?
- ¿Si es un gráfico, podrías realizar un boceto del diseño?

**Escenario 1:**

Se ha ganado un contrato de trabajo en una gran fábrica de manufactura para analizar los datos de productividad de los trabajadores para ver si puede identificar la causa de una disminución reciente en la productividad. Lo que aprendes del nuevo Gerente de Operaciones es que no importa cuántas personas adicionales contrata, el resultado es una productividad reducida. Cuando el Gerente de Operaciones fue contratado, hace seis meses, el Gerente General le dijo que la productividad se había mantenido plana durante años, y que era su trabajo aumentarlo en un 20% durante los años de recaudación. Hasta ahora ha disminuido en un 20%.

Después de escuchar este resumen del Gerente de Operaciones, una de las primeras cosas que usted decide examinar es la posible conexión entre adiciones de personal y disminuciones de productividad. Dado sus años de experiencia como analista de productividad, no le sorprende descubrir que los aumentos de personal están proporcionalmente relacionados con la disminución de la productividad. Usted sospecha que la adición de trabajadores sin cambiar nada más sobre el proceso de fabricación o las instalaciones puede haber resultado en personas simplemente entrando en el camino del otro.

Usted decidió mostrar al Gerente de Operaciones la relación de aumento de personal con la disminución de la productividad antes de tomar cualquier otro paso. Tiene estadísticas diarias de personal y productividad del último año. Tanto la plantilla como la productividad permanecieron bastante estables hasta justo después de la llegada del Gerente de Operaciones. ¿En qué forma presentará su información?

**Respuesta:**

La forma más adecuada de mostrar la información para este escenario es utilizando un gráfico, ya que lo que se desea analizar es el comportamiento de la productividad en relación al número de empleados en el tiempo, en cuanto al tipo de relación cuantitativa es posible considerar correlación y un utilizar un atributo visual para definir el tiempo, en este caso se utilizó el color para mostrar la gestión del nuevo gerente de operaciones. un posible gráfico se muestra en la figura 5

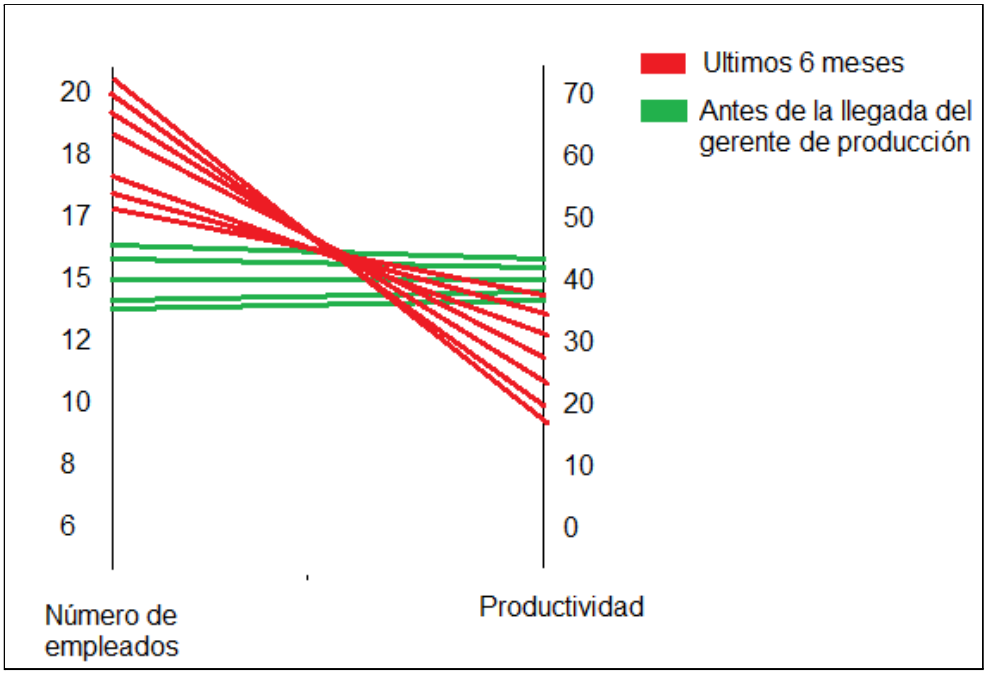


Figura 5

**Escenario 2:**

Hace seis meses desarrolló y comenzó a impartir un nuevo curso titulado "Gestión Ética". Cuando propusiste la idea inicialmente para la clase, tu director estaba muy preocupado por lo bien que sería recibido, pero tus éxitos pasados lo animaron a darte una oportunidad. Ahora que usted ha estado enseñando el curso por un tiempo y han trabajado los errores, es hora de dar a su director algunas pruebas de que él tomó la decisión correcta confiando en su juicio y capacidad.

Usted ha enseñado el curso cuatro veces durante el último mes a un total de 100 estudiantes.

Cada estudiante llenó un formulario de evaluación al final de la clase, y usted ha tabulado los resultados. En una escala de calificación de 1 a 5, con 1 representando “pérdida de tiempo” y 5 representando “excelente”, la calificación mediana para el curso es 4, y la media 4.3. Las calificaciones son excepcionales. No sólo es la calificación mediana alta, el rango de calificaciones está fuertemente agrupado alrededor de la calificación de 3, 4 y 5 con muy pocas calificaciones de 2 y ninguna de 1. Cuando se comparan las calificaciones a las que recibió por otra clase popular que usted también enseña, sus promedios eran casi iguales pero la extensión de clasificaciones para estas otras clases se distribuía de forma más amplia, indicando que no llegaba a todos los estudiantes, así como su nuevo curso.

Usted desea dar esta información a su director en una forma que él entienda de la forma más simple. En una ocasión anterior, cuando trató de comunicar las diferencias en el rango de clasificaciones entre las clases usando desviaciones estándar, se podía decir que el director

realmente no entendía cómo interpretarlas, pero estaba demasiado avergonzado para admitirlo. Esta vez usted quiere abordar la tarea diferente, sin el uso de términos estadísticos.
¿Qué forma tomará su presentación?

**Respuesta:**

Este problema se resuelve fácilmente mostrando una gráfica de barras comparando la distribución del nuevo curso, junto con el antigüo curso que era más popular (figura 6). Teniendo el gráfico de barras se puede realizar una comparativa mirando fácilmente en qué curso se tiene la mayor cantidad de aprobaciones y sus distribuciones. En este ejemplo, el mínimo caso fue de estudiantes a los que no les gusto ni les desagrado el curso, y hay una mayor cantidad de estudiantes a los que les agrado el nuevo curso. Por otro lado, ya no hay estudiantes que consideren que este curso fue una pérdida de tiempo.

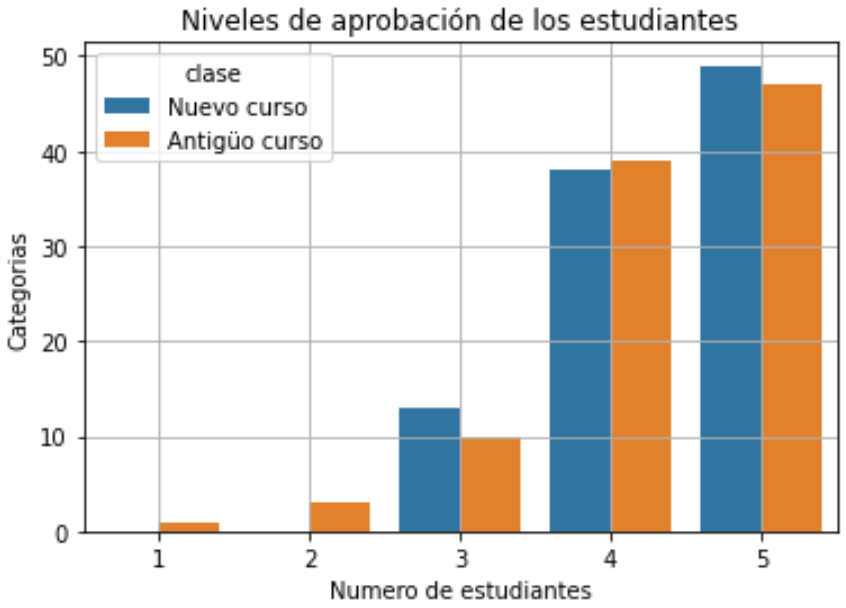


Figura 6. Distribución de la aprobación de los estudiantes.

**Ejercicio 4**

Selecciona uno de los siguientes sets de datos adjuntos a esta guía de ejercicios:

*penguins.json* : datos que describen características biológicas de pingüinos observados en diferentes zonas del planeta. Disponible también en https://vega.github.io/vega-lite/examples/data/penguins.json

*Lifeplane.csv* : datos que hablan de accidentes aéreos hasta la fecha del 21 de Septiembre del 2017, especificando detalles de los modelos de los aviones en que ocurrieron los incidentes, cuando fue introducido el modelo, cuando se retiró y cantidad de modelos en operación.

*Nobel.zip* : variedad de datos que describe características de los premios nobles de los últimos tiempos.

*station_366.json* : subset de datos de historias de trayectos en la ciclovía de New York disponible en https://www.citibikenyc.com/system-data , el subset se refiere a trayectos realizados desde y hacia la estación 366 durante Noviembre del 2011 disponible en https://raw.githubusercontent.com/vda-lab/vda-lab.github.io/master/assets/station_366.json

Explora la data representada visualmente las diferentes relaciones cuantitativas, finalmente explica que has encontrado de interesante valiéndote de visualizaciones y textos. Recuerda que a veces una visualización no unas respuestas a tus preguntas, pero una visualización si te puede revela hechos que están escondidos entre los datos y puede dar sugerencias o guías hacia donde mirar.

**Respuesta:**

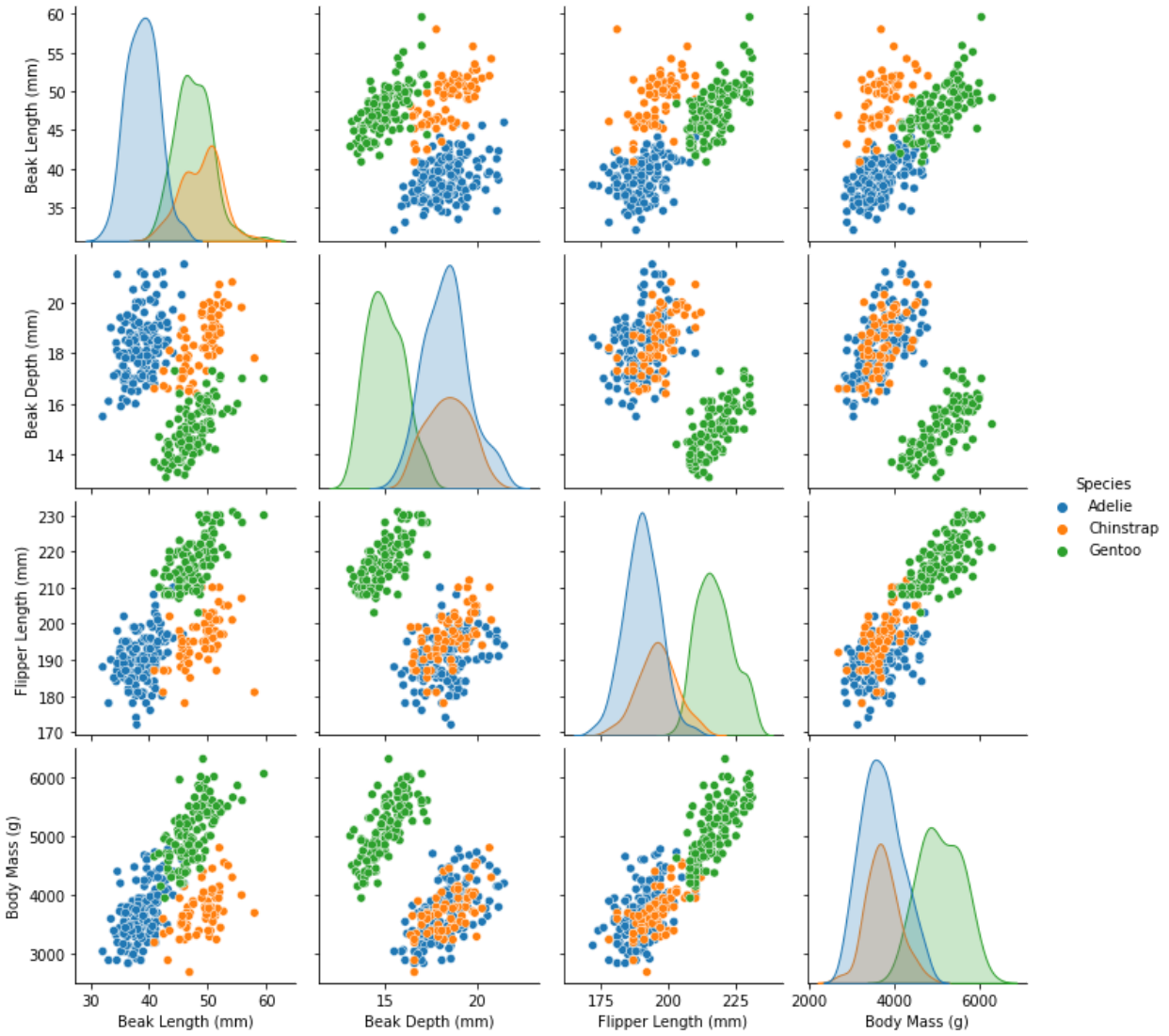


Figura 7. Pairplot características físicas de los pingüinos

El diagrama de correlación estilo pairplot observado en la figura 7, está configurado para darnos la correlación de las variables del dataset penguin, diferenciando cada punto por la isla a la que pertenece. Este diagrama es interesante desde el punto de vista del machine learning, puesto que, permite analizar el comportamiento de los datos y ver si cada una de las clases (en este caso, la especie del pingüino) puede ser diferenciable con las variables que se tienen. Por ejemplo, la variable Longitud del pico, permite diferenciar de manera excepcional entre las 3 clases de pingüino. Por otro lado, la variable profundidad del pico permite diferenciar la especie gentoo de las demás. Así mismo, de la gráfica anterior se puede concluir que las especies adelie y chinstrap, presentan características similares en cuanto a la masa del cuerpo y la profundidad del pico.

Adicionalmente, se puede ver claramente en los scatter plots, que las características se encuentran altamente correlacionadas entre sí, por ejemplo, la variable masa del cuerpo y

la longitud de la aleta presentan una alta correlación positiva, esto nos indica que entre más pesado sea un pingüino, más larga será su aleta.

Por otro lado, las variables longitud de la aleta y profundidad del pico presentan correlación negativa, pero como no existe una correlación tan marcada no puede hablarse en términos muy certeros de que al tener una aleta grande, vaya a tener una profundidad de pico pequeña, esto a la dispersión de los datos, aún así la tendencia se mantiene. Finalmente, esta gráfica puede ir muy bien apoyada de una matriz de correlación para verificar y ver de manera más numérica, los datos presentados anteriormente.

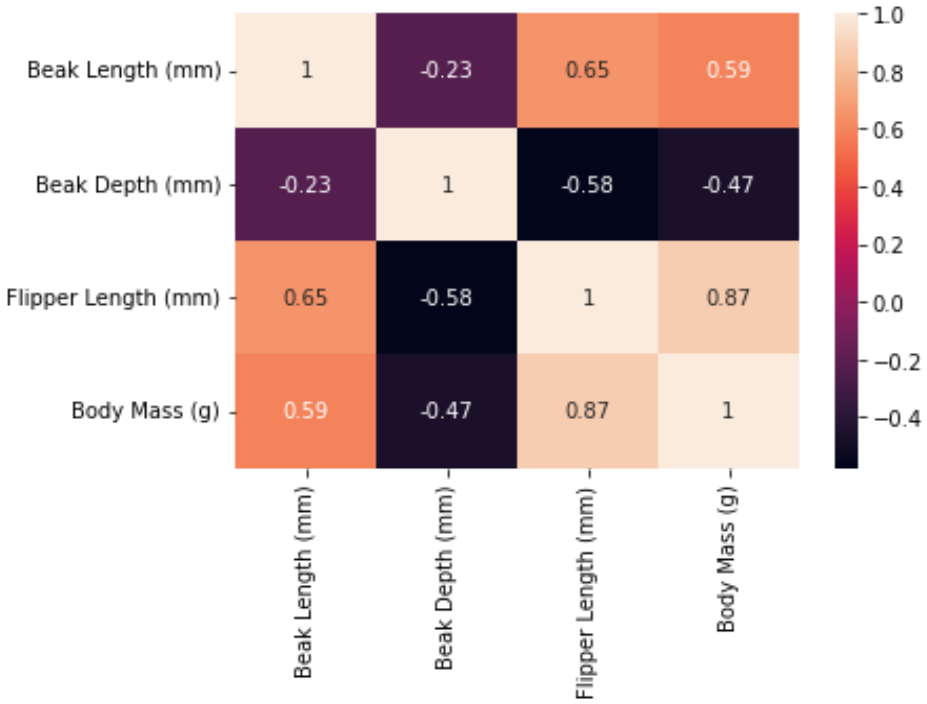


Figura 8. Matriz de correlación de las características físicas de los pingüinos.

En la figura 8 podemos apreciar varias de las observaciones que se dieron a partir de la gráfica de pairplot, pero de manera más numérica.
Por último, tenemos la gráfica de categorías paralelas (figura 9), que nos brinda información relevante sobre la distribución de los datos

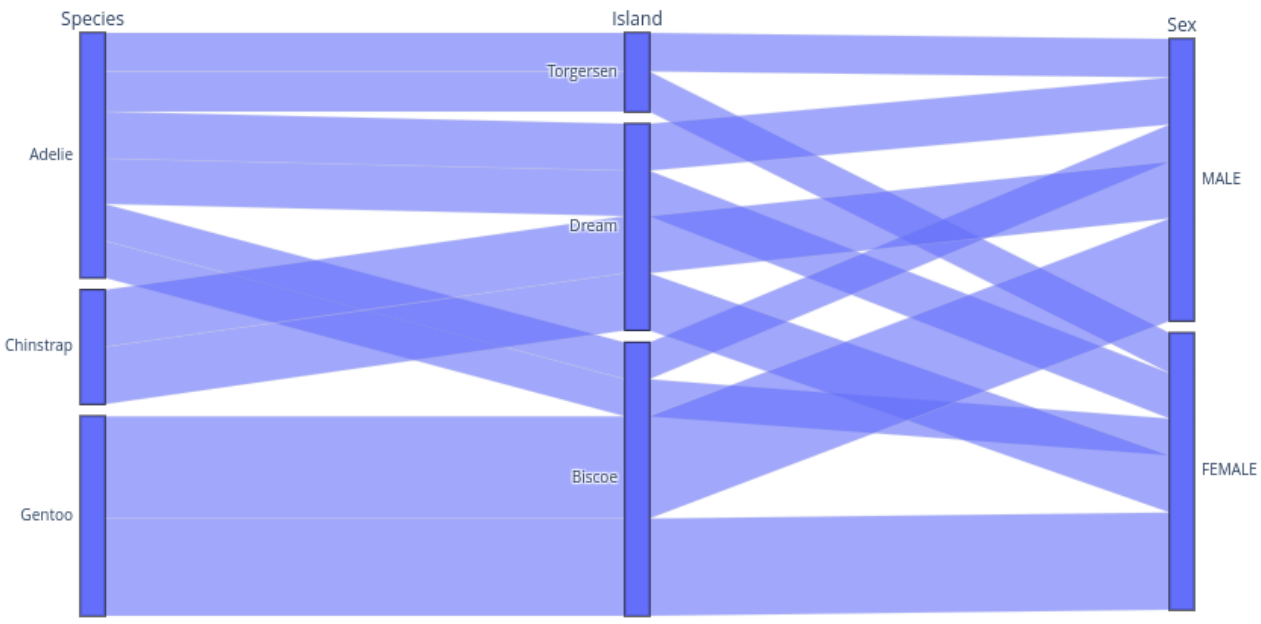


Figura 9.

La figura 9 nos permite sacar las siguientes conclusiones sobre nuestros datos:

- La especie de pingüino adelie está presente en todas las islas, lo que nos puede indicar que es la especie más común de pingüino y que ha migrado a las 3 islas.
- La isla torgersen está completamente poblada por pingüinos de la especie adele.
- La mitad de la población de pingüinos en la isla dream es de la especie chinstrap y la otra mitad es de la especie adele.
- Aproximadamente un tercio de la población de la isla Biscoe es de la especie adele y el resto es de la especie gentoo.
- La población de los pingüinos de la especie chinstrap es mucho menor a las otras especies, lo que podría indicar que su población se encuentra en riesgo de desaparecer.

**Ejercicio 5**

What is going on in this graph? https://www.nytimes.com/column/whats-going-on-in-this-graph
El sitio web del periódico *The New York* que se menciona con anterioridad te invita a mirar detenidamente unos gráficos, selecciona uno de ellos y para el mismo contestar las siguientes preguntas en el mismo orden:

1- ¿Qué notas en el gráfico? recuerda identificar la/la relación cuantitativa que está presente y qué patrón se observa.
2- ¿Qué preguntas te podrías hacer a partir de lo que notas?
3- Escribe en un párrafo la idea principal de qué está pasando en el gráfico.

**Respuesta**

Para responder a este ejercicio se seleccionó el gráfico, "*Compare the historical share of global carbon emissions" (*Figura 10*)*

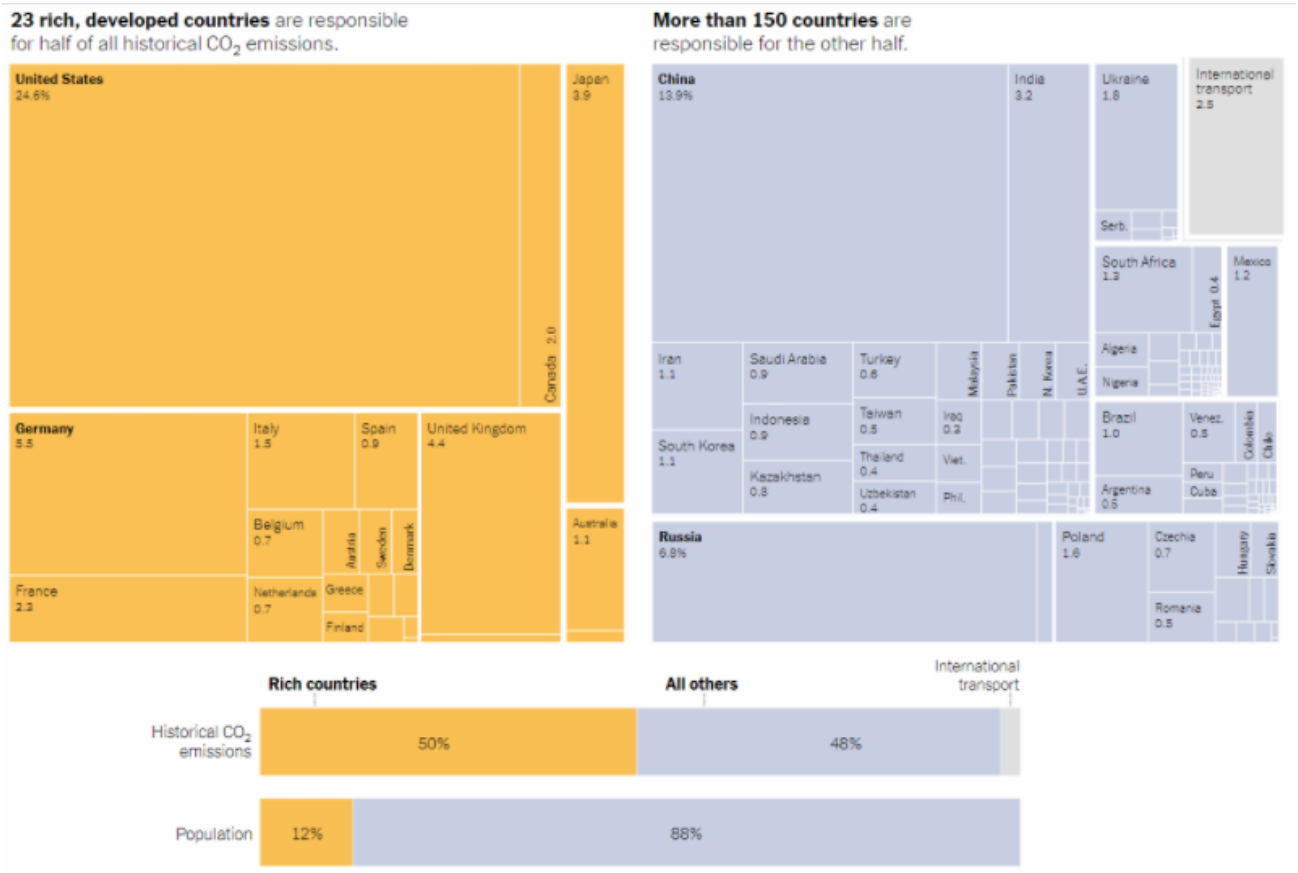


*Figura 10*

**¿Qué notas en el gráfico? Recuerda identificar la relación cuantitativa que está presente y qué patrón se observa.**

Se puede observar cuatro componentes en el gráfico que brindan distinta información acerca de los mismos datos, los dos gráficos más grandes hacen una presentación sobre las emisiones de carbono en los países ricos (cuadros de color naranja) y los otros países (cuadros de color gris). Por otro lado, las gráficas de barras inferiores muestran información resumida de

las 2 más grandes teniendo como ejes principales la emisión total de CO2 y la población. Además, se hace uso de los colores para ayudar al lector a intuir aquellos países que presentan mayor emisión de CO2 sin necesidad de revisar las cifras a detalle.

En cuanto a las relaciones cuantitativas, la primer que alcanza a vislumbrarse es una relación de jerarquía, debido a que, la información se presenta subdivida en categorías cada vez más pequeña; la primera división se hace por la riqueza de los países, la segunda de acuerdo a las regiones del planeta y la tercera por cada país dentro de la región. Así mismo, podemos encontrar la relación parte de un todo, donde al observar cada una de las categorías mencionadas con anterioridad se puede describir la distribución de las emisiones de CO2 del planeta.
Por otro lado, la disposición del grafico nos permite inferir la existencia de una correlación, mirando detalladamente las dos barras al final de la gráfica, podemos concluir que entre más riqueza presenta un país, su emisión de CO2 es mayor que en los países menos desarrollados.

**¿Qué preguntas te podrías hacer a partir de lo que notas?**

1. ¿Cuál es el valor porcentual de emisiones de CO2 por persona, industria y otras fuentes?
2. ¿Qué porcentaje de emisión de CO2 aportan las ciudades capitales de los diferentes países?
3. ¿Qué tipos de industrias generan la mayor cantidad de emisión de CO2 por país?
4. ¿Cómo se relaciona su actividad económica principal con su porcentaje de emisión de CO2?
5. ¿Los países con mayor índice de emisión de CO2 se encuentran generando planes de acción ambiental?
6. ¿Cómo se han comportado las emisiones de CO2 desde que inició la revolución industrial hasta el momento actual para los diferentes países?

**Escribe en un párrafo la idea principal de qué está pasando en el gráfico.**

Se ha demostrado a través de los años que existe una fuerte correlación entre el desarrollo económico de un país y sus emisiones de CO2. La visualización de los datos históricos de emisión de CO2 pretende dejar en evidencia el impacto ambiental que sucede como efecto secundario al avance económico de un país. La distribución de la gráfica nos permite identificar de manera rápida los países con mayor emisión de CO2. El gráfico hace la distinción entre 23 países considerados más desarrollados, demostrando que los países ricos con tan solo el 12% de la población mundial son responsables del 50% de emisión de CO2. De acuerdo a ello se puede intuir que el alto porcentaje de contaminación se atribuye a la industrialización, mientras que, en países como china que históricamente han sufrido de sobrepoblación, su alto índice de contaminación se atribuye a las personas y sus actividades cotidianas.

**Ejercicio 6**

Trabajas en el departamento de Atención al Cliente, posees la data de los tickets (solicitudes) recibidas y procesadas, en Mayo dos empleados se fueron de la empresa. Tu colega quiere presentar este gráfico al jefe para contarle de la situación. Analiza la data, has un listado de los problemas que encuentras, haciendo uso de la técnica de data-link ratio y los Principios del Diseño Analítico reconstruye el mismo tal que cuente una historia de una manera clara y placentera. Los datos se encuentran en *ticket_trend.json* adjunta a esta guía de ejercicios.

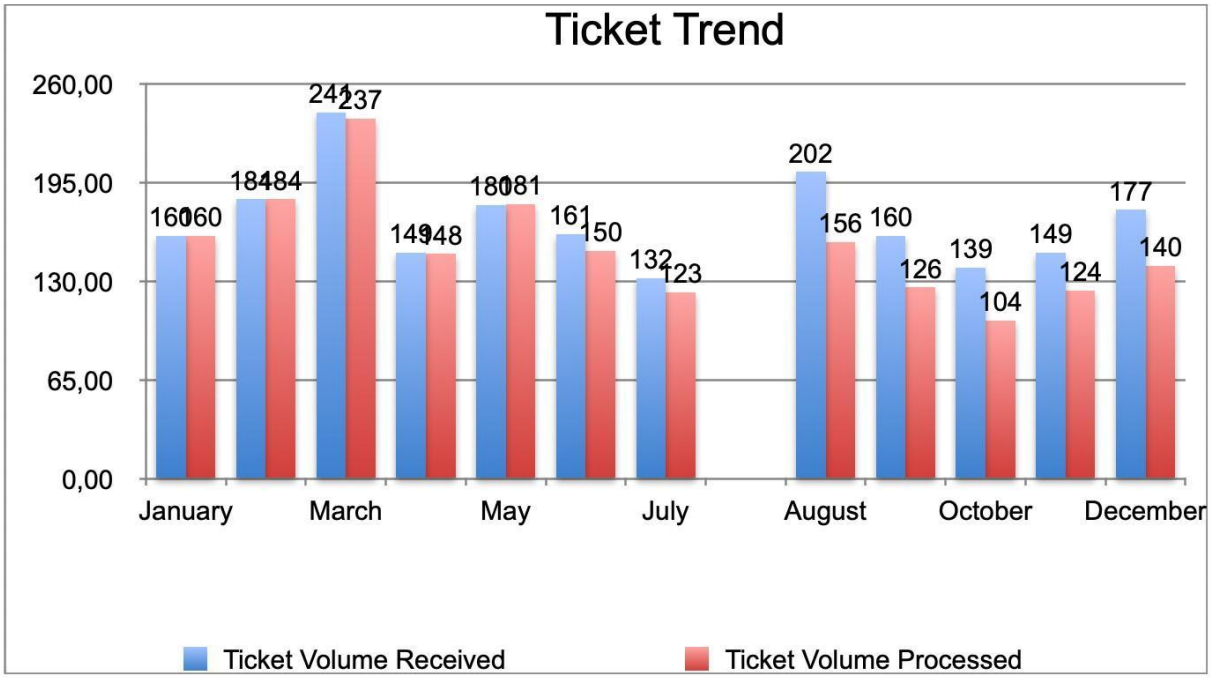


Figura 11

**Respuesta:**

Luego de identificar los datos mostrados en la gráfica anterior (figura 11), se puede evidenciar que existe un aumento progresivo de los tickets desatendidos a partir del mes de mayo, según el propósito del gráfico, el cual es notificarle al jefe la renuncia de dos empleados y cómo ha afectado el rendimiento del trabajo, se presentan algunas de las fallas:

- Para el propósito del gráfico no es necesario mostrar los datos de los tickets recibidos y los tickets procesados.
- No es necesario mostrar las cantidades en las barras para los tickets recibidos y procesados

- ❖ Existe un espacio entre el mes de julio y agosto, el cual hace pensar que falta algún mes o se quisiera agrupar los meses en dos grupos.
- ❖ El ancho de las barras es insuficiente y los números se encuentran superpuestos.
- ❖ No existen etiquetas para el eje x y el eje y

En el siguiente gráfico (figura 12) se muestra una relación entre la cantidad de tickets recibidos y la cantidad de tickets desatendidos, se normalizaron estos valores con el objetivo de comparar cada una de las señales y evidenciar un aumento en los tickets desatendidos desde mayo, evitando que el lector imagine que ocurre un aumento en las ventas.

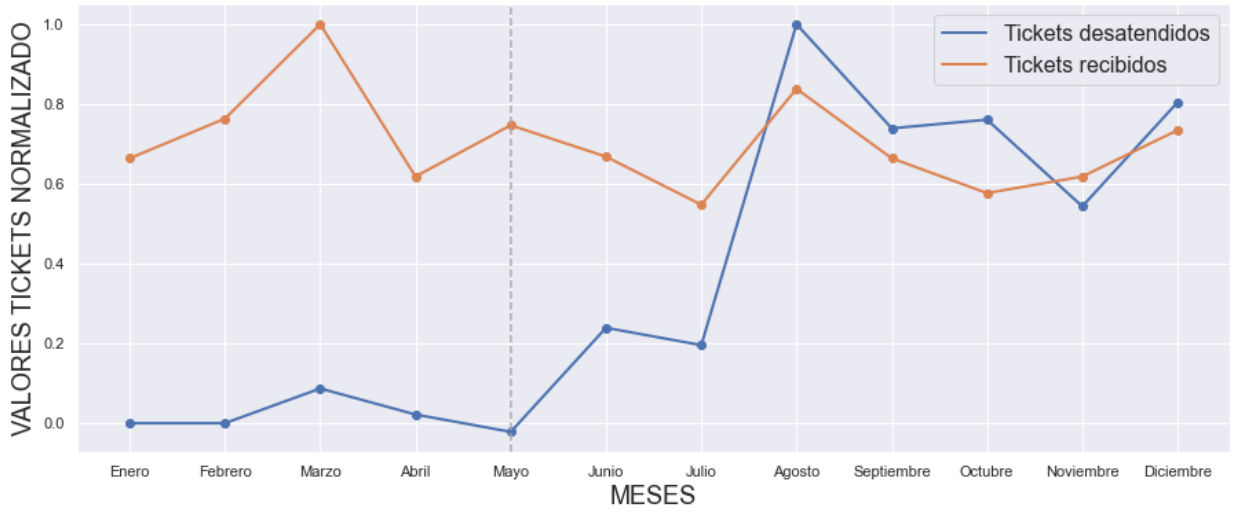


Figura 12

**Ejercicio 7**

Nuestro departamento de Ciencias de la Universidad ha realizado un Programa para promover la Ciencia en estudiantes de bachillerato. Ahora desean presentar los resultados de las encuestas realizadas antes y después al comité de financiamiento para que se pueda volver a repetir el programa. Tus colegas quieren presentar este gráfico que se muestra a continuación. Analiza la data. Haciendo uso de la técnica de data-link ratio y los Principios del Diseño Analítico cómo crees que presentarías los resultados para comunicar el éxito del programa realizado. Los datos se encuentran en *encuestas.json* adjunta a esta guía de ejercicios

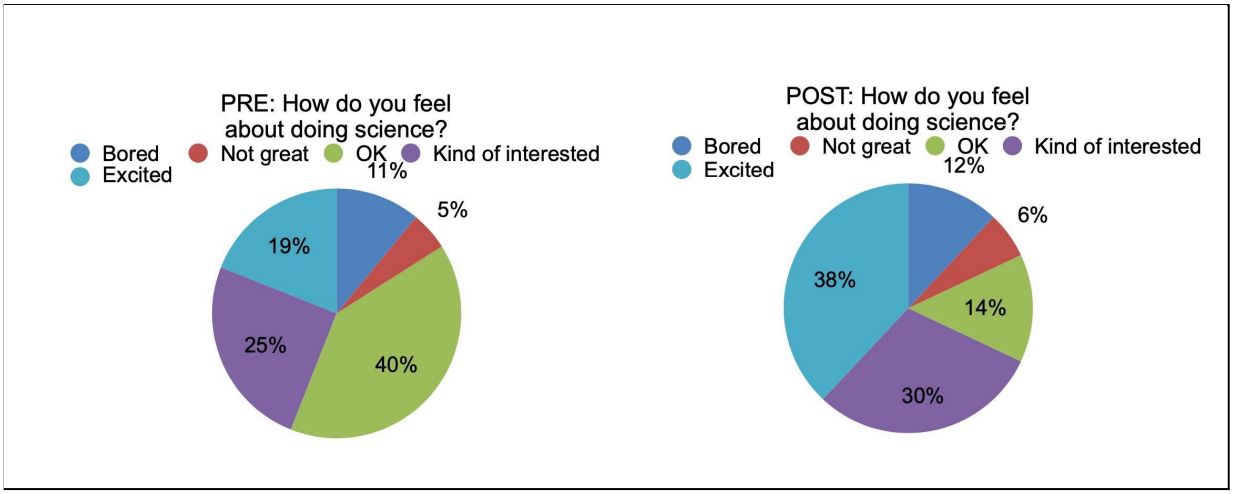


Figura 13

**Respuesta:**

El gráfico anterior (figura 13), permite identificar cual es el porcentaje de estudiantes que presentan cierta postura con respecto a la ciencia, calificandolo en: aburrido, no tan genial, una postura neutral “ok”, interesante y emocionante. El objetivo del gráfico es comparar estas posturas antes y después de aplicar un programa para promover la ciencia, con este tipo de gráfico se dificulta o requiere de gran esfuerzo conocer el resultado final.

En la figura 14 se muestra un gráfico simplificado que me permite conocer de forma rápida el porcentaje de estudiantes que presentan una postura negativa, neutral o positiva, a su vez que compara el antes y después de haber aplicado el programa para promover la ciencia. Nótese cómo los estudiantes que presentan una postura negativa se mantiene, pero, se puede evidenciar una reducción significativa de los estudiantes con una posición neutral pasado a tener una opinión positiva con respecto a la ciencia, evidenciando de forma clara la efectividad del programa.

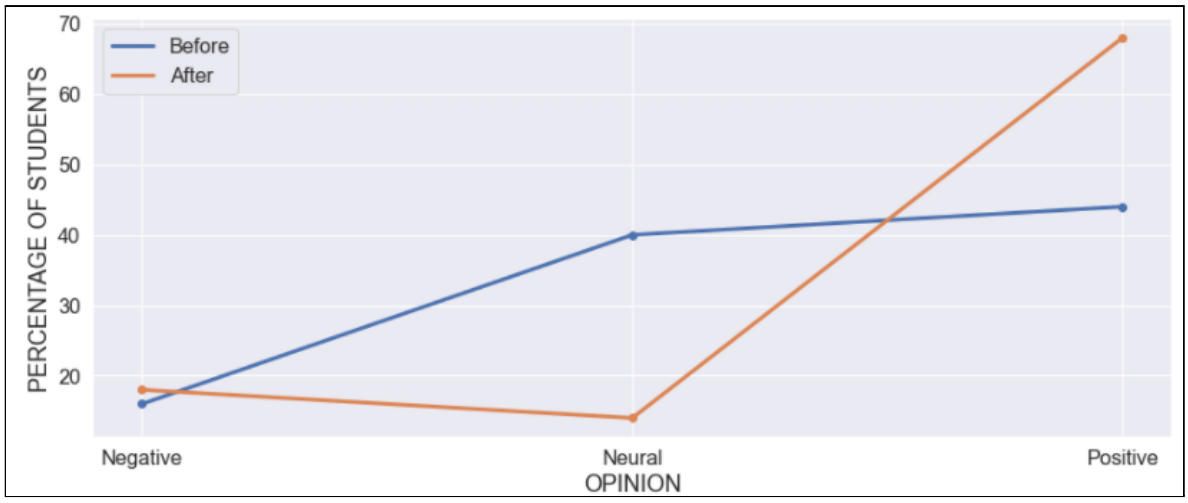


Figura 14.

**Ejercicio 8**

Make Over Monday (https://www.makeovermonday.co.uk) es un proyecto participativo para aprender a como visualizar datos de una manera efectiva, cada semana los organizadores publican una visualización encontrada en alguna publicación e invitan a que los participantes a que analicen los datos y rediseñen la visualización. En este ejercicio usted seleccionara uno de los set de datos publicados por MakeOverMonday (https://www.makeovermonday.co.uk/data/) estudie la visualización original y la fuente de donde ha salido, acorde a lo aprendido en clases describa que le parece que esta bien y que le parece que esta mal y por ultimo construya su propia versión de la visualización.

**Respuesta:**

La selección para este problema se encuentra en el siguiente link https://data.world/makeovermonday/2018-w-1-u-s-per-capita-consumption-of-poultry-livestock.

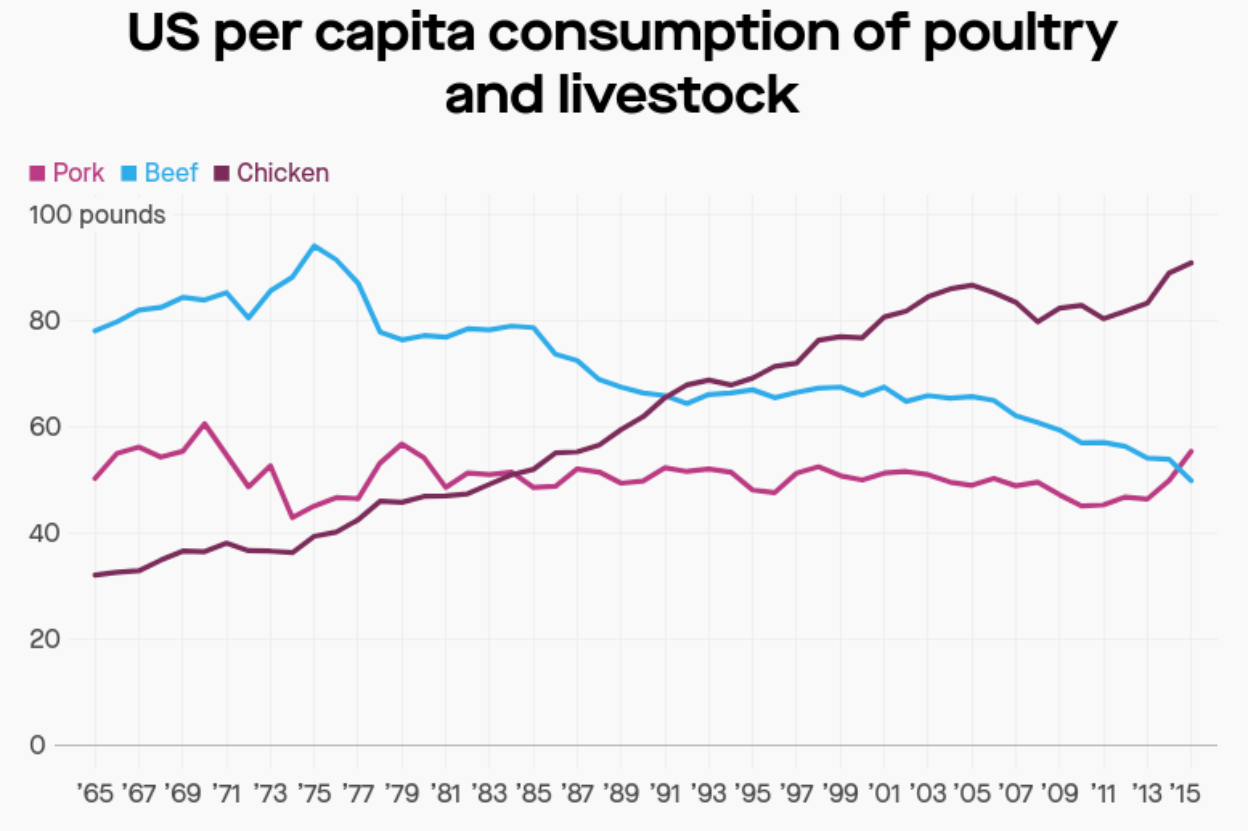


Figura 15. Gráfica de la actividad.

Lo malo:

1. La información presentada no está bien descrita por el título. Habla de ganado y pone el cerdo y la vaca en categorías diferentes.
2. El eje x no se entiende a qué variable hace referencia.
3. La información presentada debería tener el mismo cero, ya que el trabajo corresponde con una comparación de consumo.
4. La escala en el eje y está mal proporcionada.
5. El color del pollo y del cerdo debería ser más diferenciable.
6. No se muestra algún punto donde empiezan a ocurrir cambios importantes en los hábitos de consumo de la población.
7. Hay información del dataset que no se utiliza en la gráfica y que según el objetivo de visualización, debería estar presente (consumo de pavo y pollo de engorde).

Lo bueno:

1. Es el gráfico indicado para el objetivo de visualización (análisis de consumo de carne de ave de corral y carne de ganado).
2. Tiene etiquetas que permiten diferenciar cada serie de tiempo.

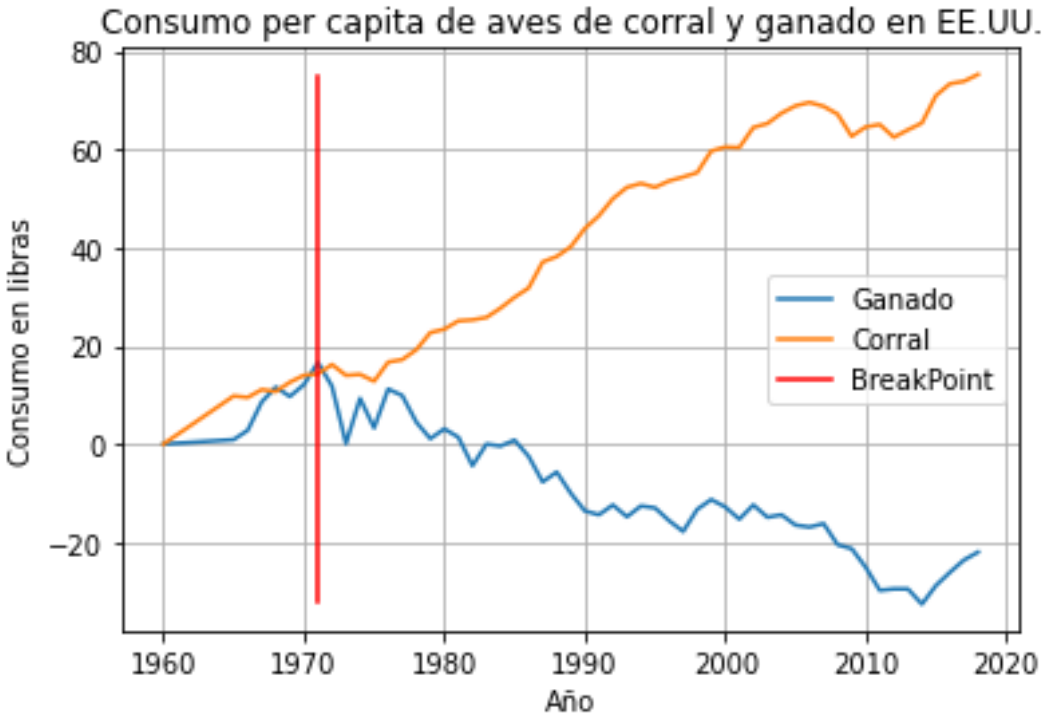


Figura 16. Gráfica corregida con las observaciones de la primera gráfica.