**Universidad Peruana de Ciencias Aplicadas**

**Escuela de Postgrado**

**Maestría en Data Science**



**Trabajo Final**

**Revision de cuatro casos de uso Grupo 5**

**Curso:** Gestion de Datos

**Profesor:** Oscar Ramos

**Presentado por:** Ortega Baca, Aldo Javier

González Victoria, Jorge Wilfredo

Berríos Aguilar, Fabrizio Franco Antonio

Aybar Coronel, Jorge Luis

**Lima – 2022**

Indices

[Introducción 3](#_Toc102142356)

[Primer Caso Estudio de evolución de la Mortalidad de Neonatos y la importancia del lavado de manos antes de atender a los pacientes en el XIX 4](#_Toc102142357)

[Segundo Caso: Análisis del consumo alcohol en Rusia 4](#_Toc102142358)

[Tercer Caso: Control de Pesos en la Crianza de Aves 4](#_Toc102142359)

[Cuarto Caso Análisis de Outliers de precios en autos usados de la marca BMW 5](#_Toc102142360)

[Metodología y Resultados. 5](#_Toc102142361)

[Primer Caso Estudio de evolución de la Mortalidad de Neonatos y la importancia del lavado de manos antes de atender a los pacientes en el XIX 7](#_Toc102142362)

[Segundo Caso: Análisis del consumo alcohol en Rusia 11](#_Toc102142363)

[Tercer Caso: Control de Pesos en la Crianza de Aves 27](#_Toc102142364)

[Cuarto Caso Análisis de Outliers de precios en autos usados de la marca BMW 35](#_Toc102142365)

[Conclusiones y Recomendaciones 42](#_Toc102142366)

[Bibliografía 43](#_Toc102142367)

[Anexos 43](#_Toc102142368)

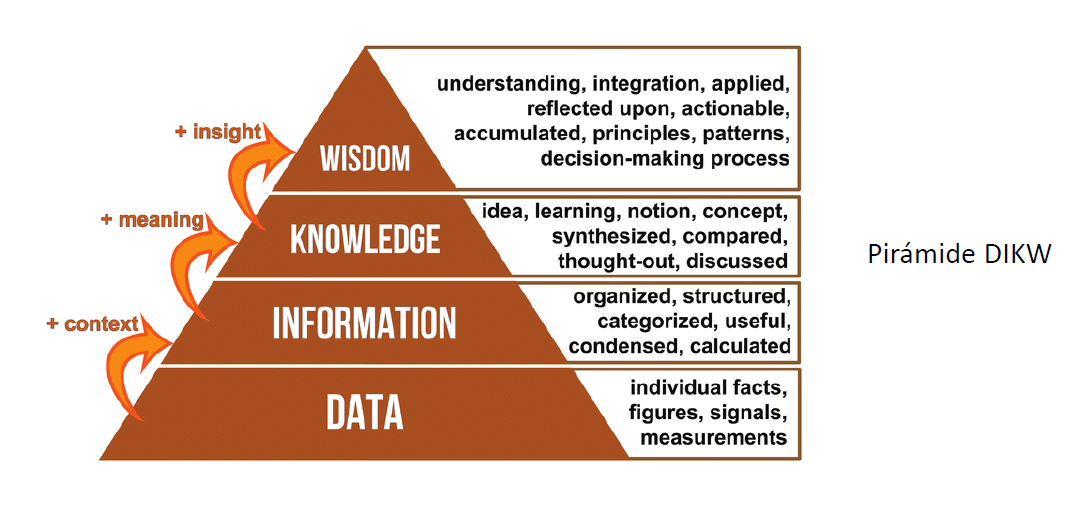
# Introducción

La gestión de datos a lo largo del tiempo ha tenido una constante evolución en conformidad a como se han desarrollado los diferentes avances tecnológicos que permiten almacenar, administrar, transmitir, preparar, etiquetar, limpiar la data para así poder preparar modelos que permitan diseñar y entregar modelos estadísticos de apoyo a la decisión basado en hechos facticos soportados por información de valor.

Como bien se sabe la data por sí sola no tiene valor si es que no se aplica un contexto el cual a través de un procesamiento se convierte en información.

Dentro de algunas herramientas y marcos de trabajo muy comúnmente utilizadas podemos mencionar lo que serían utilizados en el presente trabajo de estudio

* DAMA DMBOK
* CRISP-DM
* IBM
* SaS
* Rstudio
* Shiny App
* Python
* Collab
* Atlas Mongo DB
* AWS, GCP, Azure
* AWS MySQL EC2 instance
* AWS RDS
* GITHUB



En el presente documento se expondrá los pasos para evaluar 4 casos de uso en los cuales se aplicaran las diferentes buenas prácticas y esquemas de análisis y métodos de visualización vistas en el curso que permitirán obtener un nivel de entendimiento de la dinámica de los mismos como parte de las actividades de comprensión de los datos y que podrían ser maximizadas luego con otras técnicas de depuración, preparación y entrenamiento de datos que serán contemplados en cursos posteriores. Claro esta este curso se enfoca en explorar los diferentes esquemas de visualización de los datos.

## Primer Caso Estudio de evolución de la Mortalidad de Neonatos y la importancia del lavado de manos antes de atender a los pacientes en el XIX

El doctor Semmelweis  a mediados de 1840 observo que un 10% de los recién nacidos dos clínicas 1 y 2 fallecían por fiebre relacionada a infecciones y él recomendó que los doctores debían lavarse las manos antes de atender a los pacientes pues pensaba que por esas razones las infecciones proliferaban durante el parto. Claro está eso es un hecho factico en el XXI la importancia de las condiciones asépticas, pero en ese contexto del XIX no había forma de validar esta afirmación del doctor tenga influencia positiva en la reducción de muertes en neonatos durante el parto.

Por estas razones utilizando la data que registraba las muertes de neonatos anual y mensual registradas desde 1841 a 1846.

## Segundo Caso: Análisis del consumo alcohol en Rusia

Este caso de uso se utilizan la data de venta histórica de una cadena de tiendas en Rusia para diferentes productos descritos en ingles debido al contenido de la data: wine, beer, vodka, champagne and brandy

La compañía ejecuto una campana exitosa en la región de Saint Petersburg y la idea es poder replicarla en las 10 regiones que tengan comportamiento o hábitos de compra similares porque el área de marketing no tiene recursos para implementar campanas en todas las regiones.

En este caso se utilizaran diversos esquemas de visualización, gestión de datos perdidos, análisis de valores atípicos que permitan soportar la decisión de donde convendría aplicar la campaña de marketing

## Tercer Caso: Control de Pesos en la Crianza de Aves

El contexto de este caso está en la industria avícola de la empresa peruana Grupo Santa Elena S.A. que tiene como principal actividad la crianza de aves de corral, específicamente pollos engorde. El proceso de crianza de pollos busca lograr el máximo de kilogramos de carne de pollo y para ello debe cuidar sus parámetros productivos como la mortalidad, la nutrición y el peso corporal del ave; todos ellos obedecen a estándares dados por la genética del ave, sistema de crianza y sexo.

## Cuarto Caso Análisis de Outliers de precios en autos usados de la marca BMW

# Metodología y Resultados.

Para todos los casos expuestos en el siguiente documento se utilizar un servicio en MySQL instalado en una instancia de AWS EC2 estableciendo la conexión pertinente al entorno de desarrollo y visualización de resultados RStudio.

Específicamente, en el tercer caso de estudio se utilizara el servicio Atlas de Mongo DB con conexión al servicio de Google Collab en donde se realiza el análisis y visualización de los esquemas correspondientes

El esquema utilizado para conectar Rstudio con la instancia MySQL EC2 de AWS se muestra en la siguiente imagen.

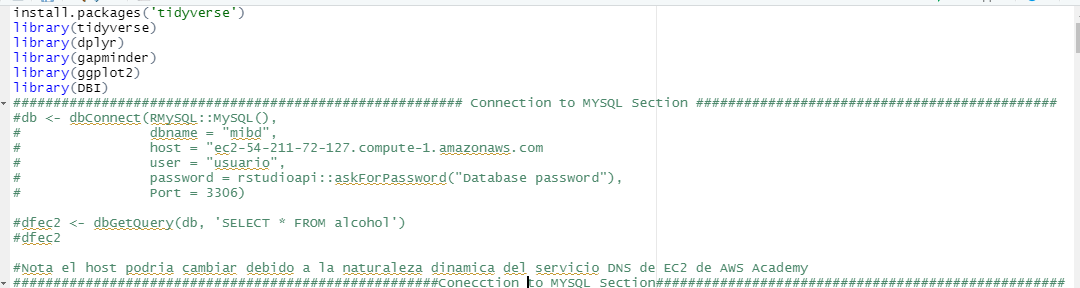


Ilustración 1 : Codigo de Conexion hacia MYSQL/EC2 desde Rtudio

Asimismo, en la siguiente imagen se demuestra la creación de las tablas en la base de datos midb en MySQL que serán utilizados en todos los casos excepto el tercero.

Las tablas creadas y ser utilizadas son las siguientes:

* Monthly.
* Yearly
* Alcohol
* BMW

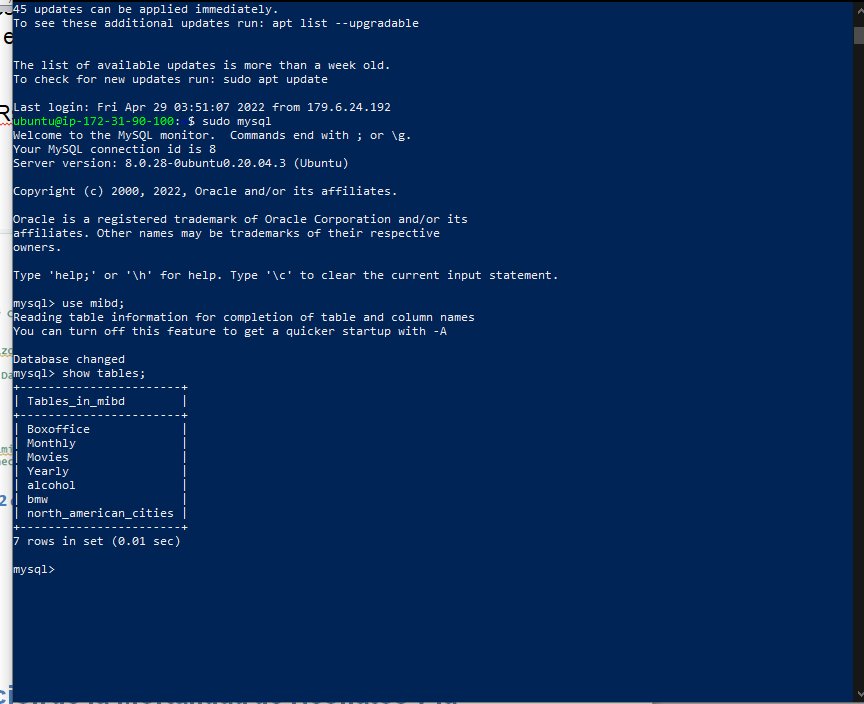


Ilustración 2 : Validación MYSQL Database on AWS EC2 Service

En el tercer caso de uso se utilizara la Bases de Datos BD\_Granjas y la colección pesos de mongo DB como se observa en la siguiente ilustración

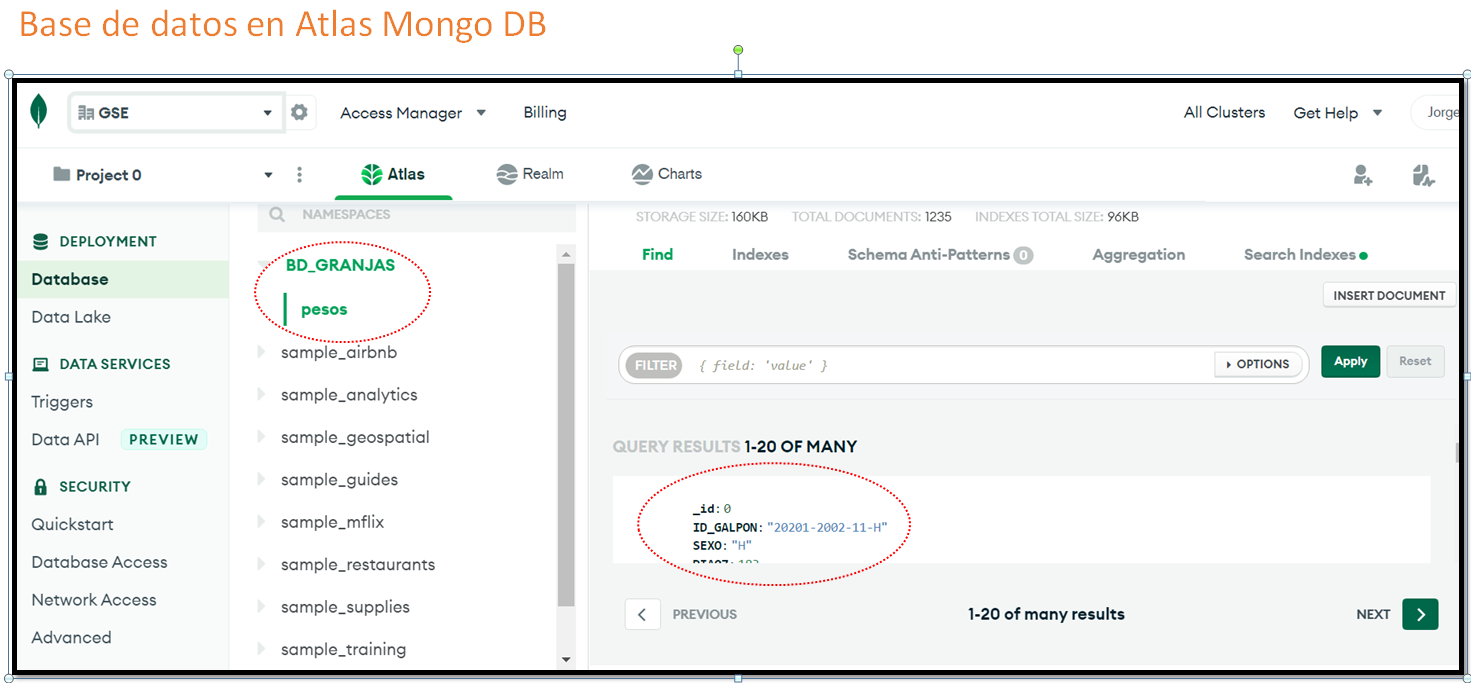


Ilustración 3: Servicio Atlas Mongo DB para el Tercer Caso

## Primer Caso Estudio de evolución de la Mortalidad de Neonatos y la importancia del lavado de manos antes de atender a los pacientes en el XIX

Para este caso de estudio los datos que serán utilizados en el modelo de evaluación visual pertenecen a las tablas Yearly and Monthly del entorno MySQL/EC2 como se muestra en la siguiente figura

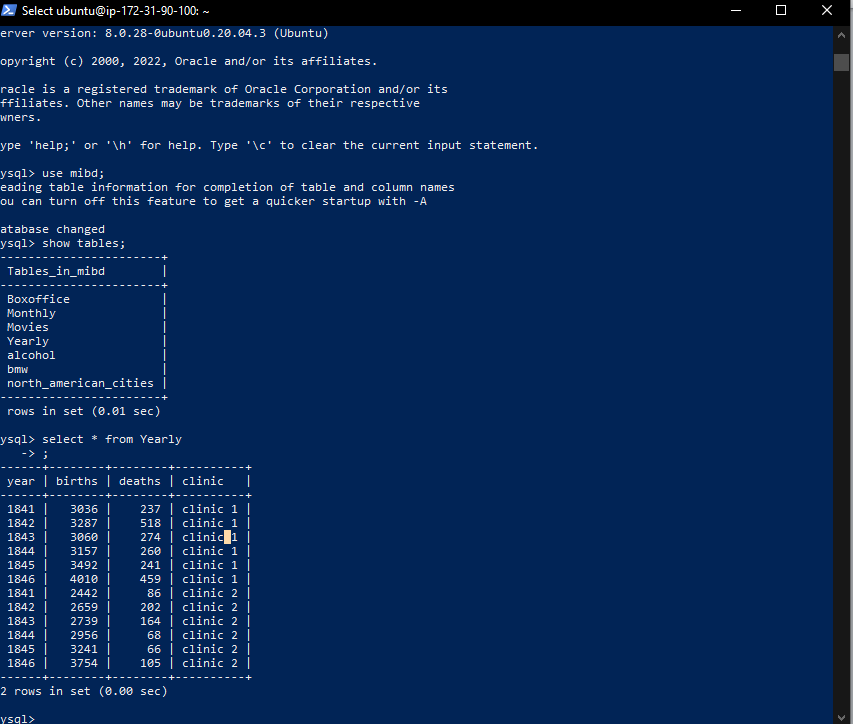


Ilustración 4: Bases de datos Yearly

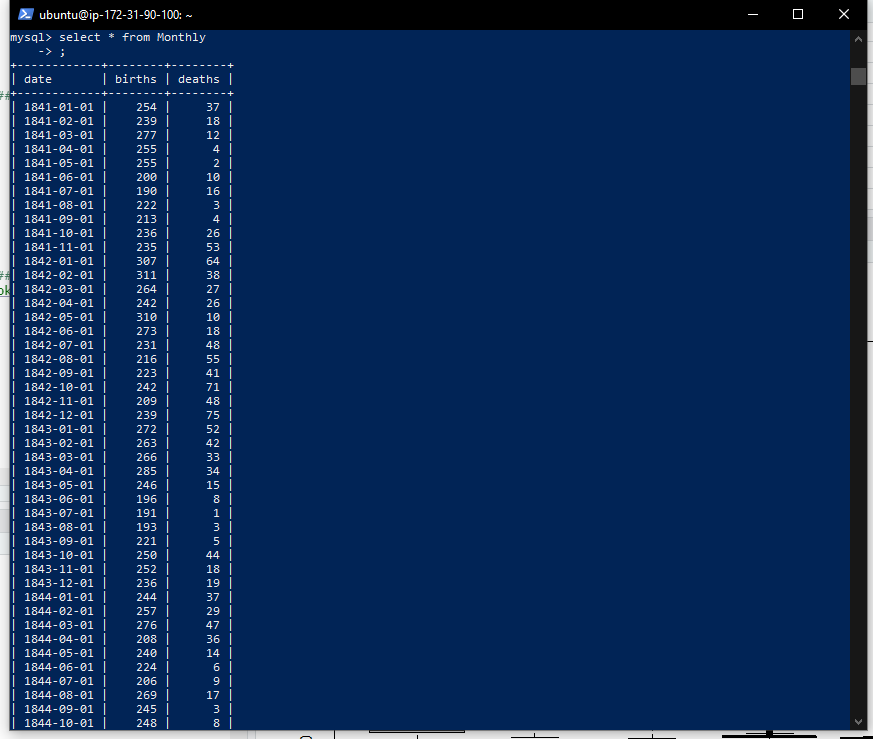
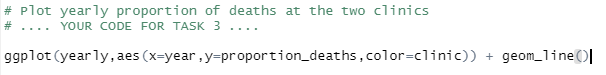


Ilustración 5: Bases de Datos Monthly

Utilizando la base datos Yearly bajo la utilizacion de las librias Tydeverse y ggplot2 se puede desarrollar la visualizacion de la evolucion de muertes por anioo en cada una de las dos clinicas 1 y 2



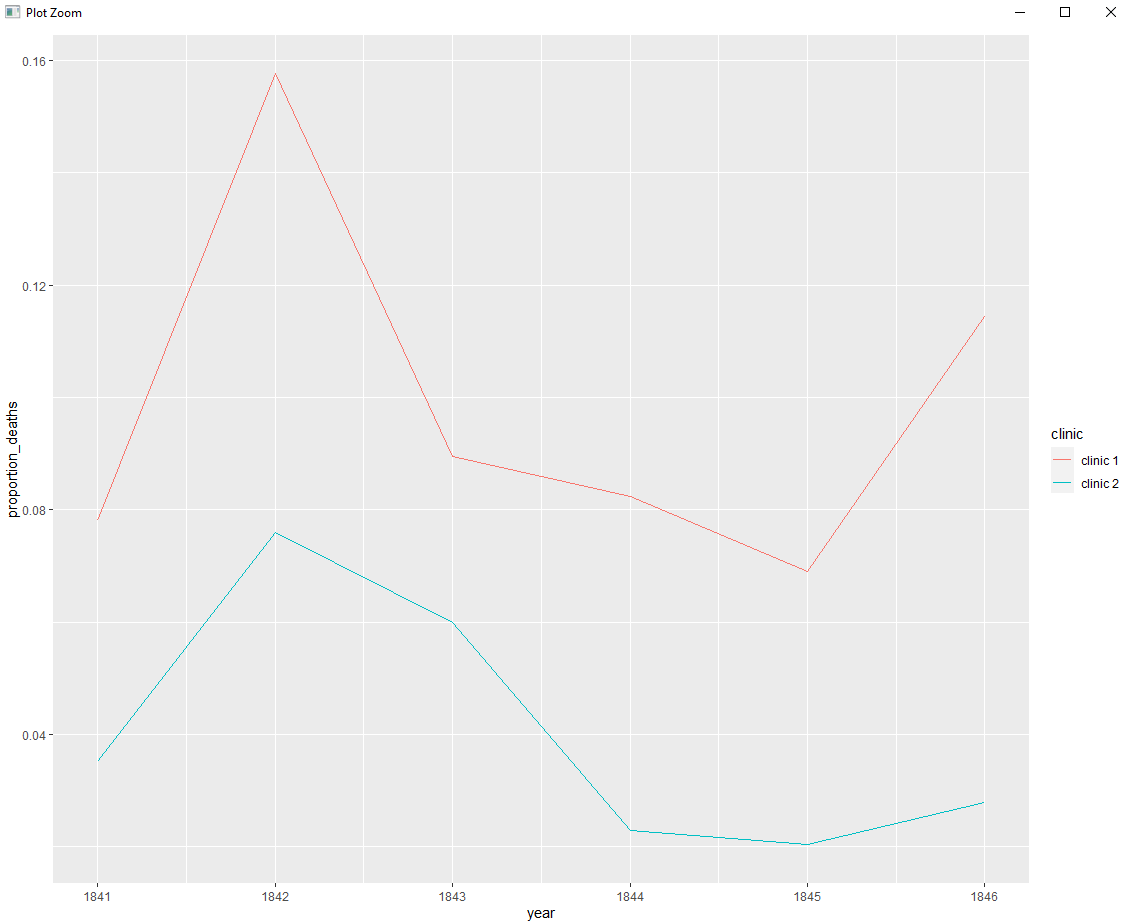
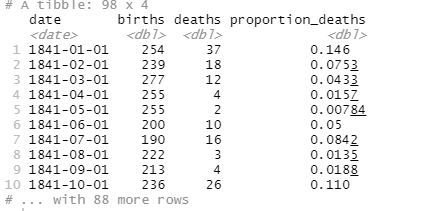
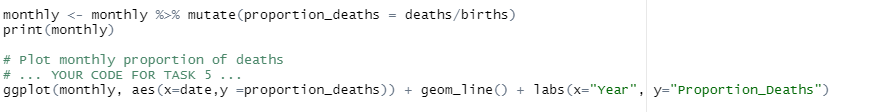


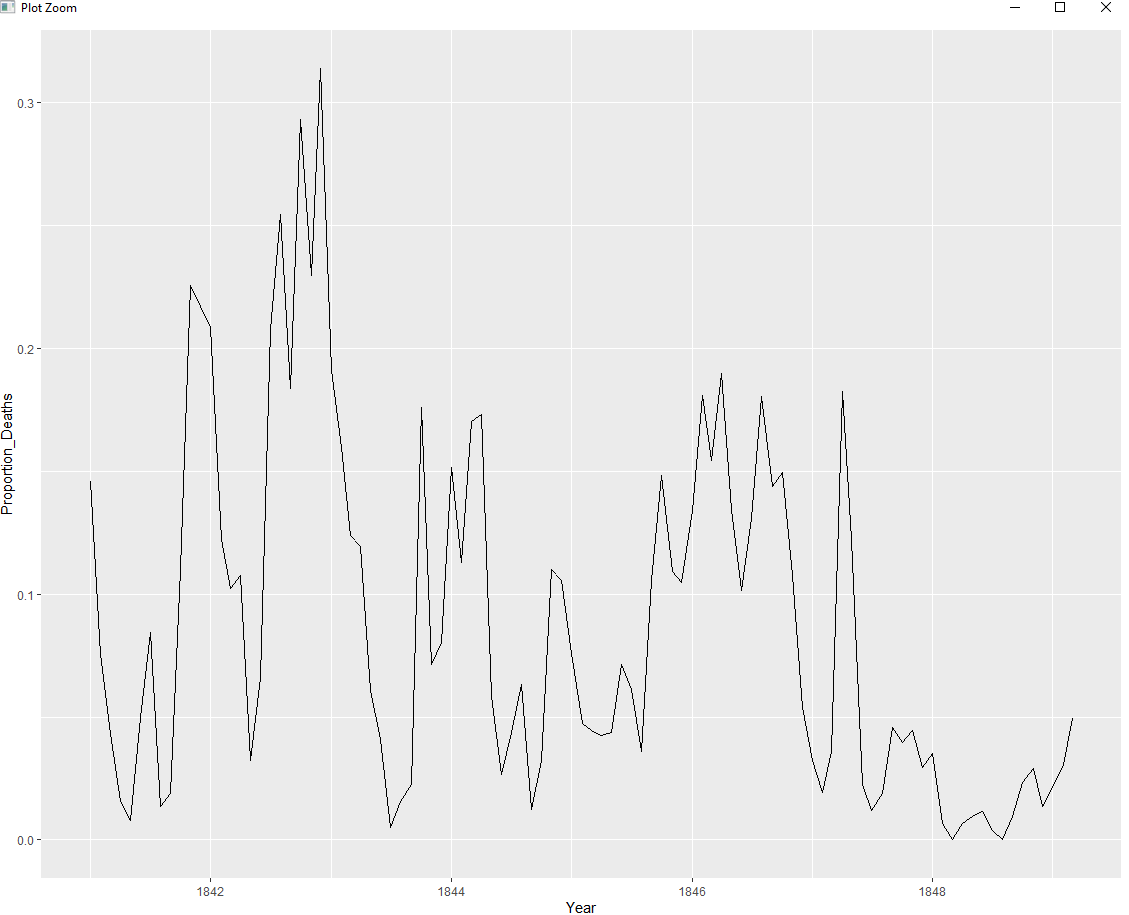
Ilustración 6 Muertes Neonatos por clinica

Luego utilizando la fuente de datos Monthly de MySQL EC2 podemos calcular la proporcion de muertes por mes durante los anios 1841 a 1846 utilizando el metodo mutate de tidyverse.

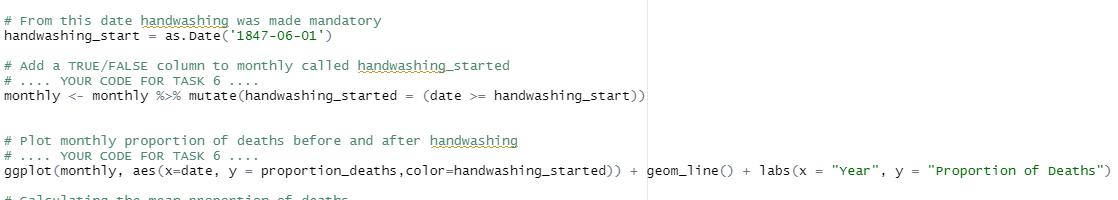


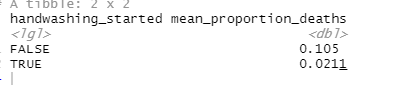


Por consiguiente se puede obtener la evolucion de la proporcion de muertes por mes desde 1841 a 1846



El Doctor Semmelweis hizo su recomendación y aseveración el 01-06-1847, por consiguiente luego de implementar esta política del lavado de manos en las dos clínicas de estudio antes de atender a las parteras, se puede visualizar una importante reducción de la proporción de las muertes de neonatos.





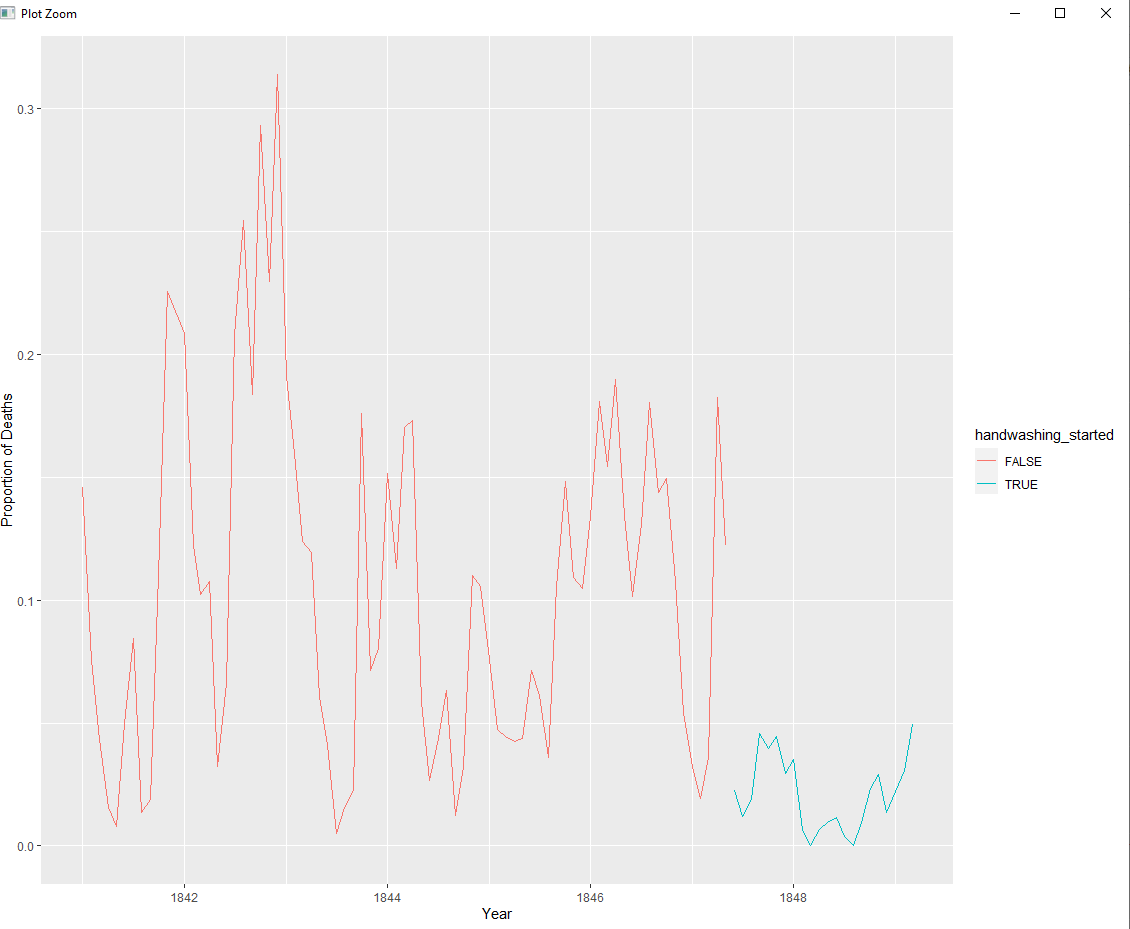
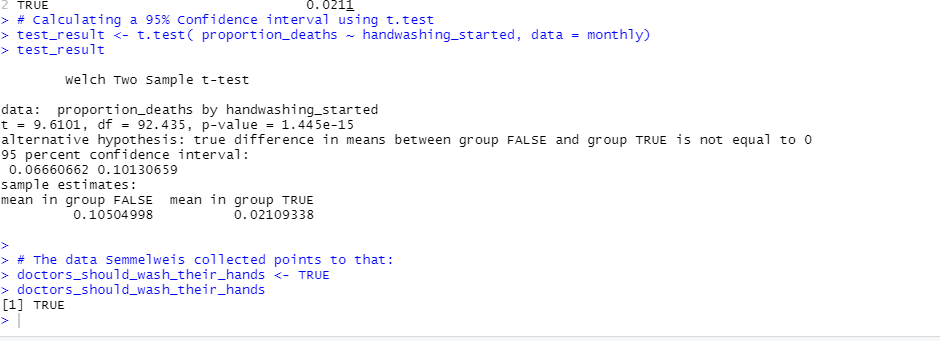


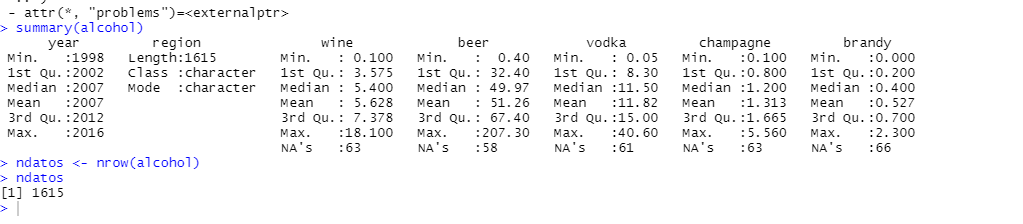
Ilustración 7: Evolución de muertes antes y después del lavado de manos

Por estas razones, podemos afirmar que las recomendaciones del Doctor permitió recudir la proporción de muertes en neonatos y dicha recomendación esta soportada por el análisis de los datos pues ser redujo la proporción de mortandad de 10.5% a 2.1%



## Segundo Caso: Análisis del consumo alcohol en Rusia

En el desarrollo de este caso de uso se utilizara la base de dato alcohol alojada en MySQL EC2 AWS instancia en conformidad a lo que se muestra en la siguiente ilustración. Asimismo, el dataset tiene 1615 rows y 7 columns



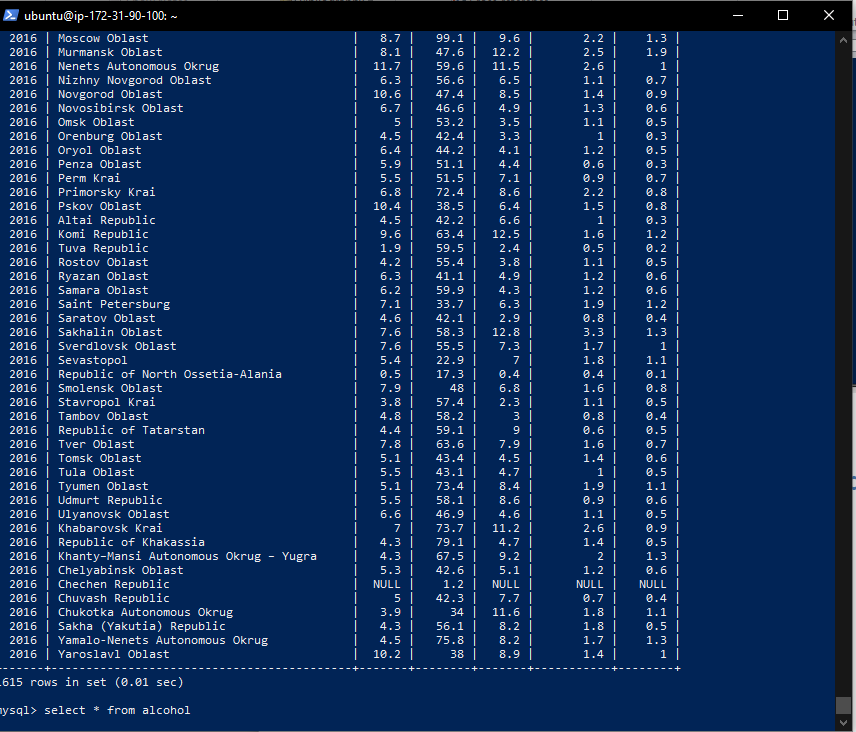
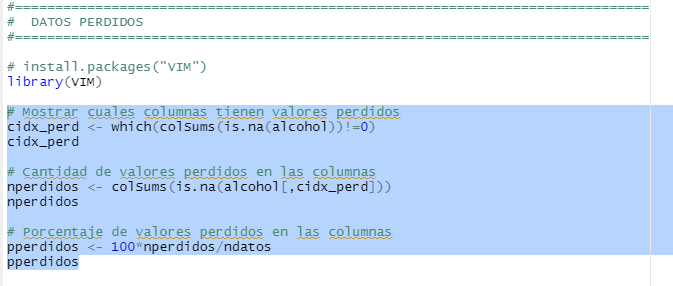


Ilustración 8 AWS Query a la DB Alcohol

Al explorar el conjunto de datos, se puede observar que existe datos perdidos sobre las ventas de los productos alcohólicos en ciertas regiones y años según la información resumen de la imagen adjunta



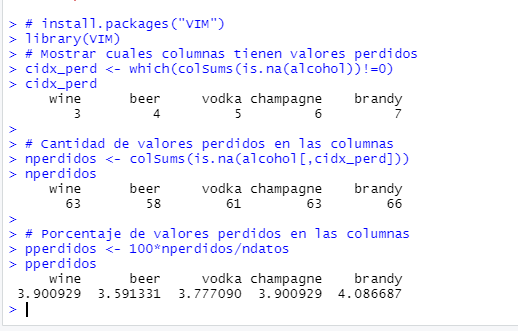


Ilustración 9: Proporcion de Datos Perdidos

Luego de explorar la data relacionada utilizando un esquema de visualización agregada se observa que la proporción de datos perdidos es menor igual a 0.04 en el peor de los casos por tanto se decide simplemente no considerar esos datos faltantes alineados a las buenas prácticas para su tratamiento vistos en el desarrollo del curso.

• 1 % de datos faltantes: trivial (el método de imputación no tiene mayor impacto)

• 1 a 10 % de datos faltantes: manejable (requiere un método “simple”)

• 10 a 20 % de datos faltantes: requiere métodos sofisticados (puede requerir método “propio”)

• Más del 20 % de datos faltantes: interpretación perjudicial (ya se perdió “demasiado”)

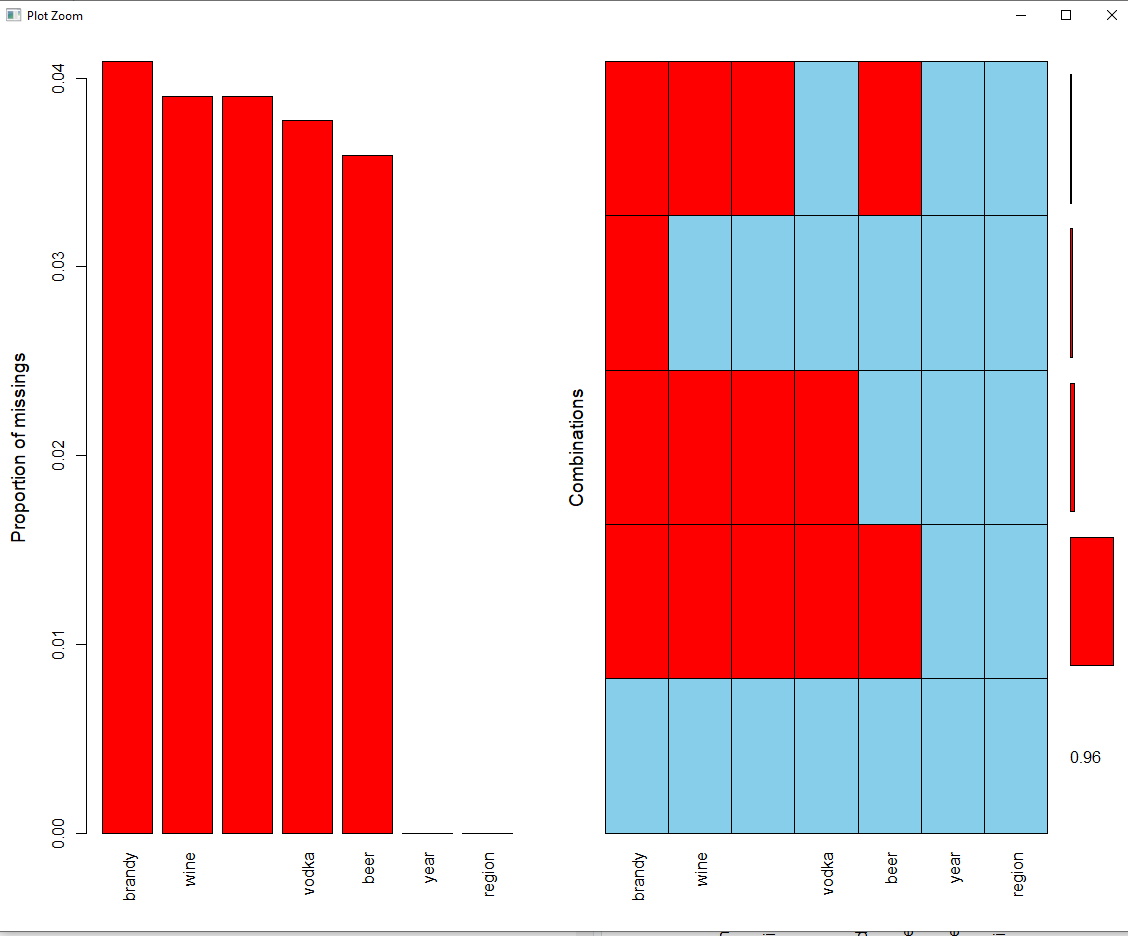


Ilustración 10: Proporcion de Datos Perdidos

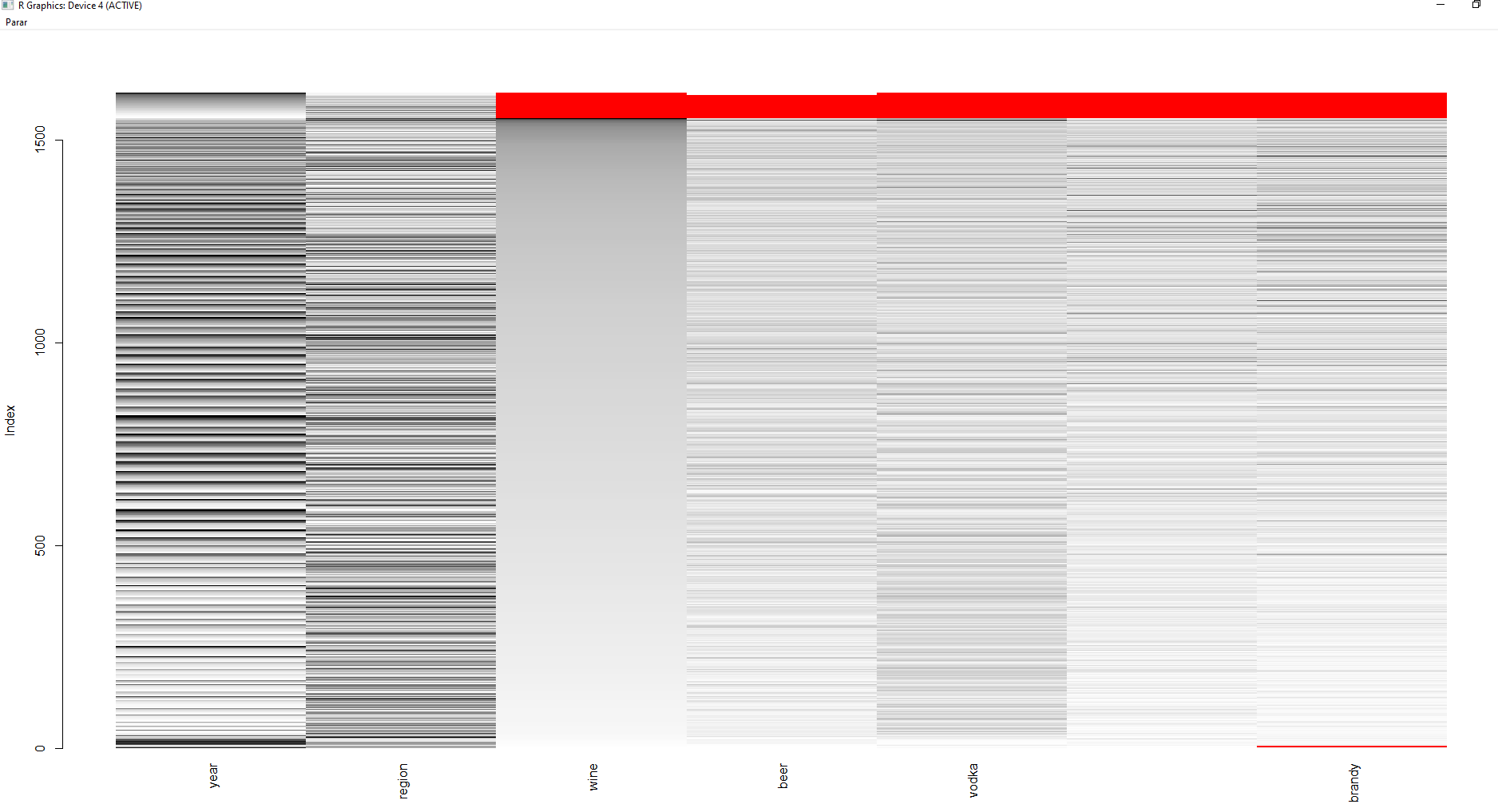
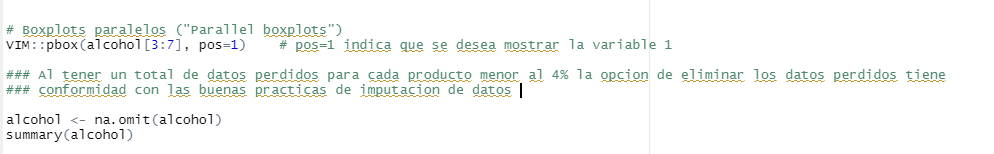
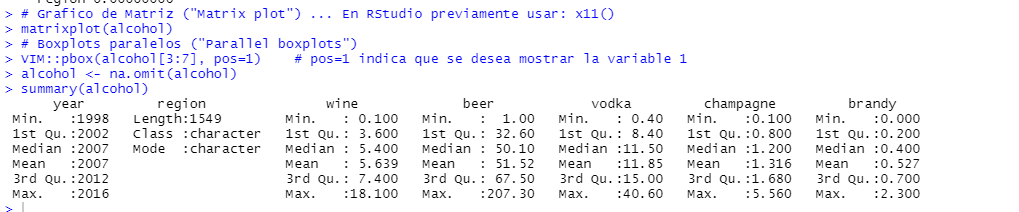


Ilustración 11: Patrones de Datos Perdidos

Por tanto se decide ejecutar el código para depurar la data eliminando los registros en Null o NA





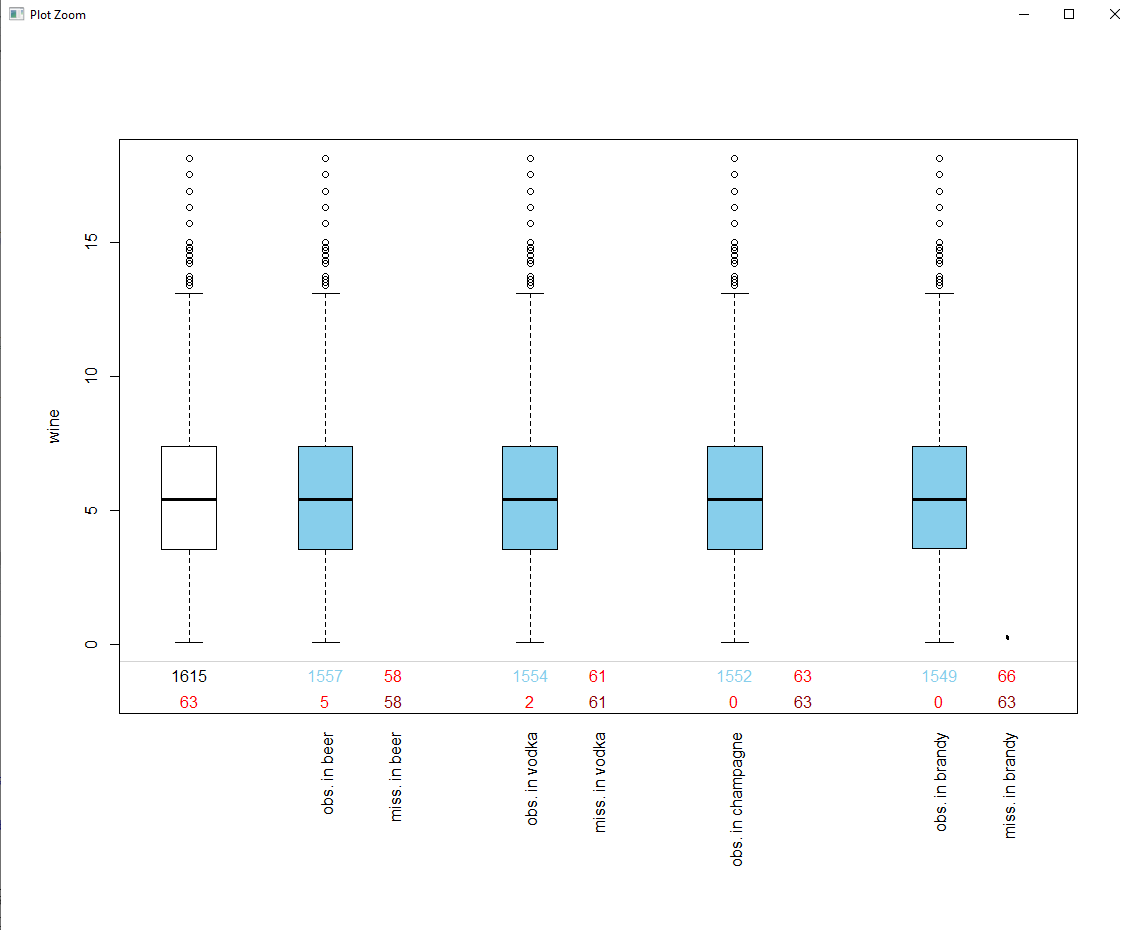


Ilustración 12: Datos Perdidos verus Observados

Al ser la ciudad de San Petersburg la ciudad patrón de ventas en donde se aplicó exitosamente la campaña de marketing, conviene visualizar cuales son las tendencias de ventas en esta región especifica utilizando la librería ggpubr de Rstudio



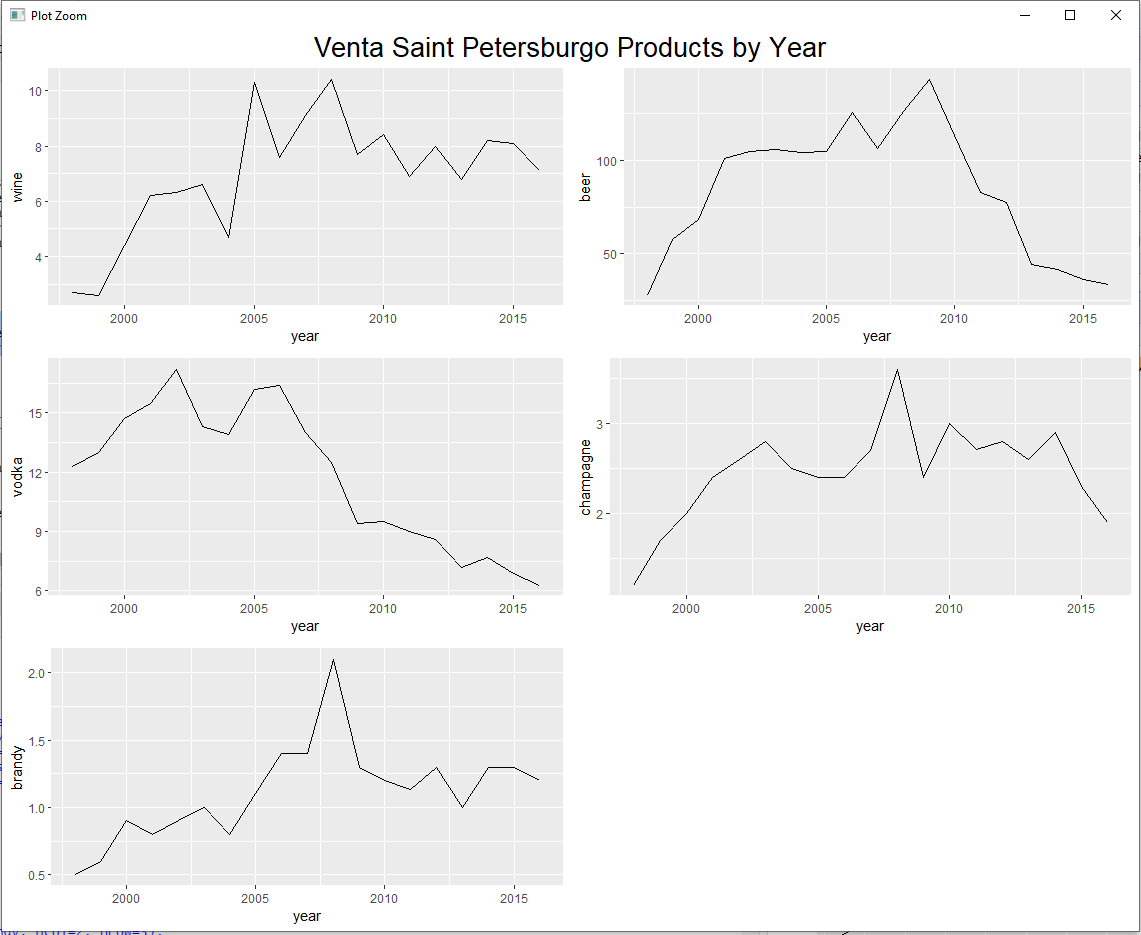


Ilustración 13: Saint Petersburg Produc Sales by Year Evolution

Luego, conviene evaluar la venta a lo largo de los años de cada producto: wine, beer, vodka, champagne, brandy identificando la tendencia anual y los TOP N regiones con mayor venta en cada línea de producto con el objetivo de identificar regiones top año en Rusia.

Se elaboró vistas interactivas en Shiny App para poder visualizar la evolución de las ventas por año

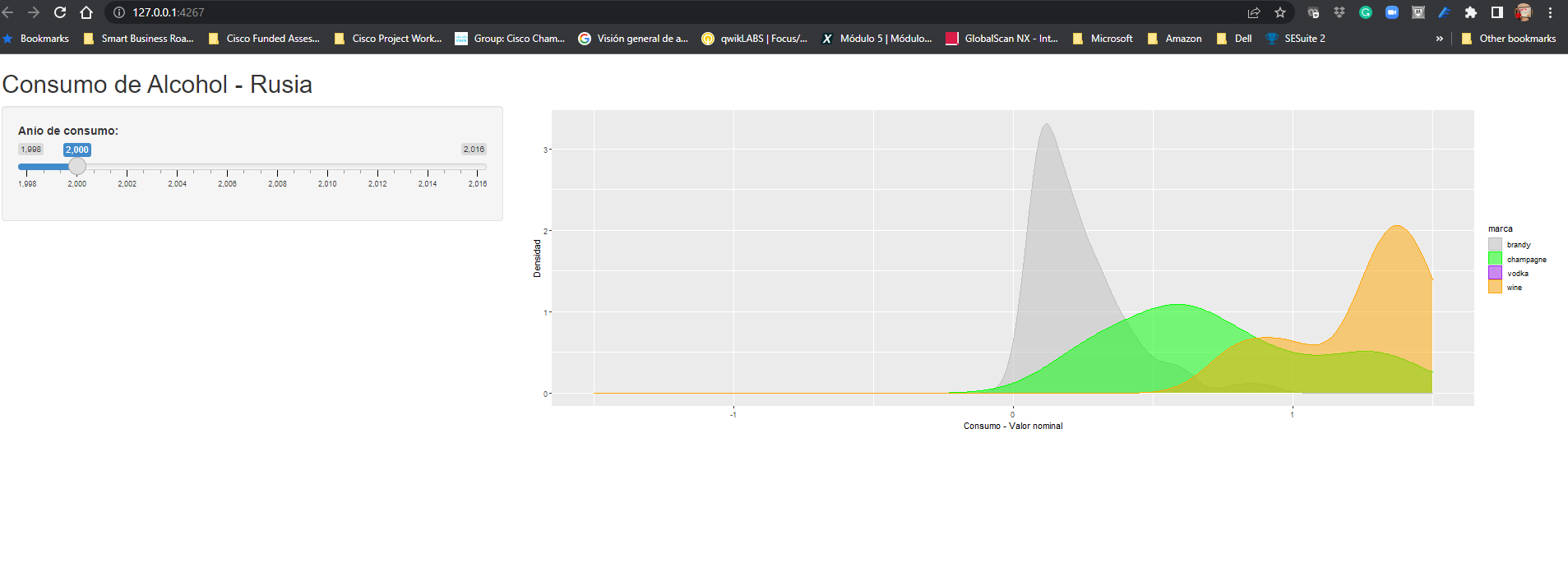


Ilustración 14: Shiny App View

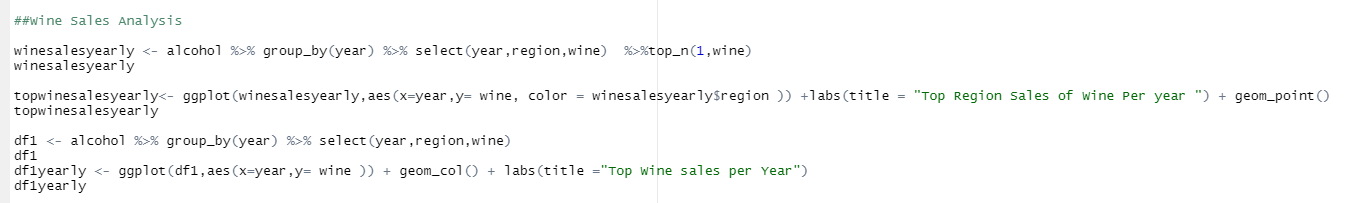


Ilustración 15: Codigo Wine

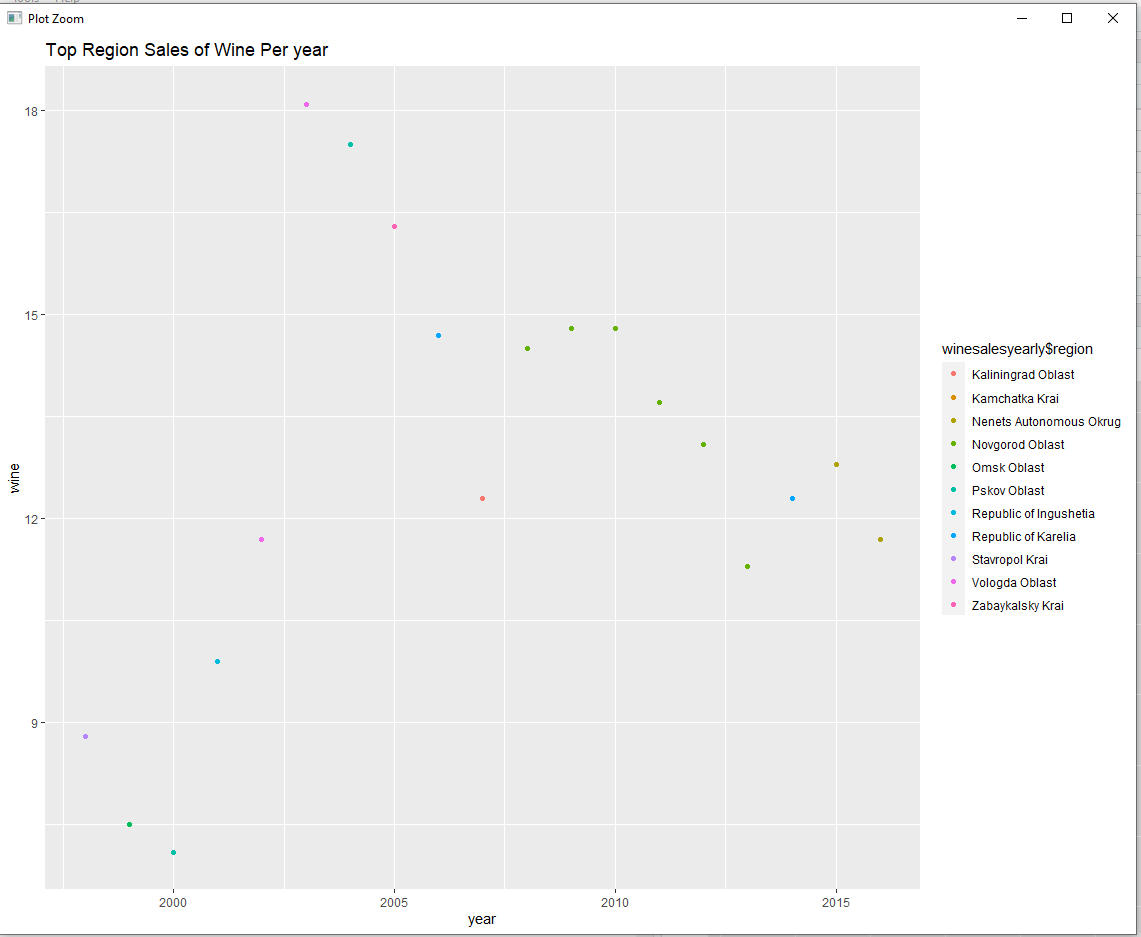


Ilustración 16: Top Wine Region Sales By Year

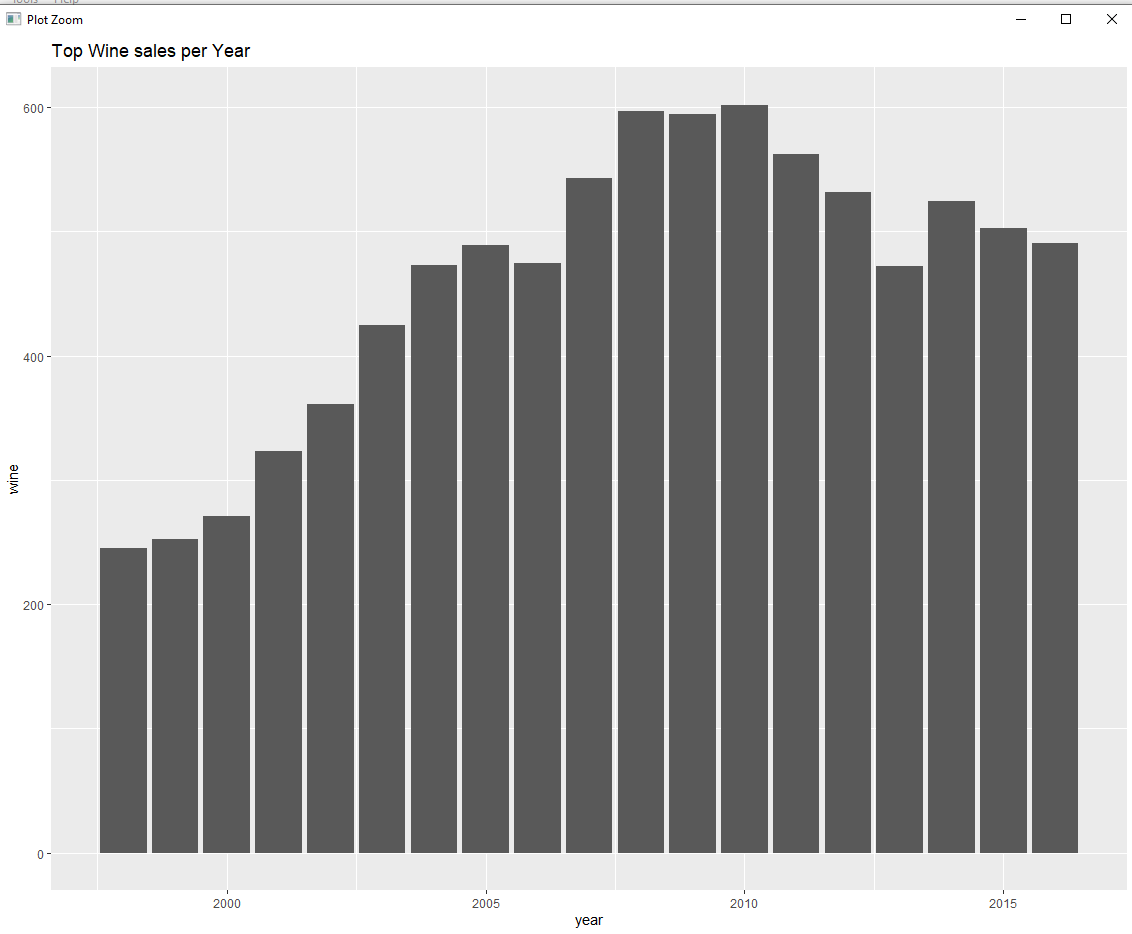


Ilustración 17: Wine Sales Evolution by Year

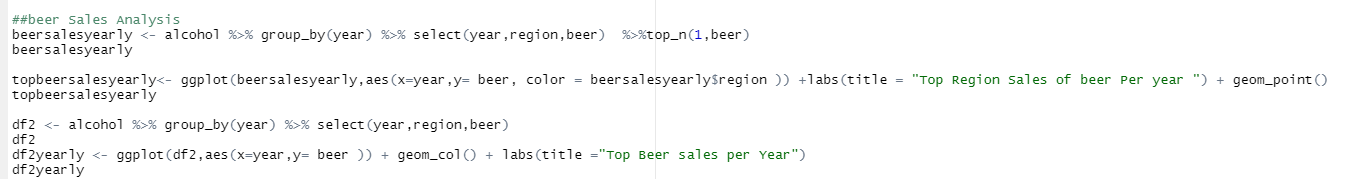


Ilustración 18: Codigo R Beer

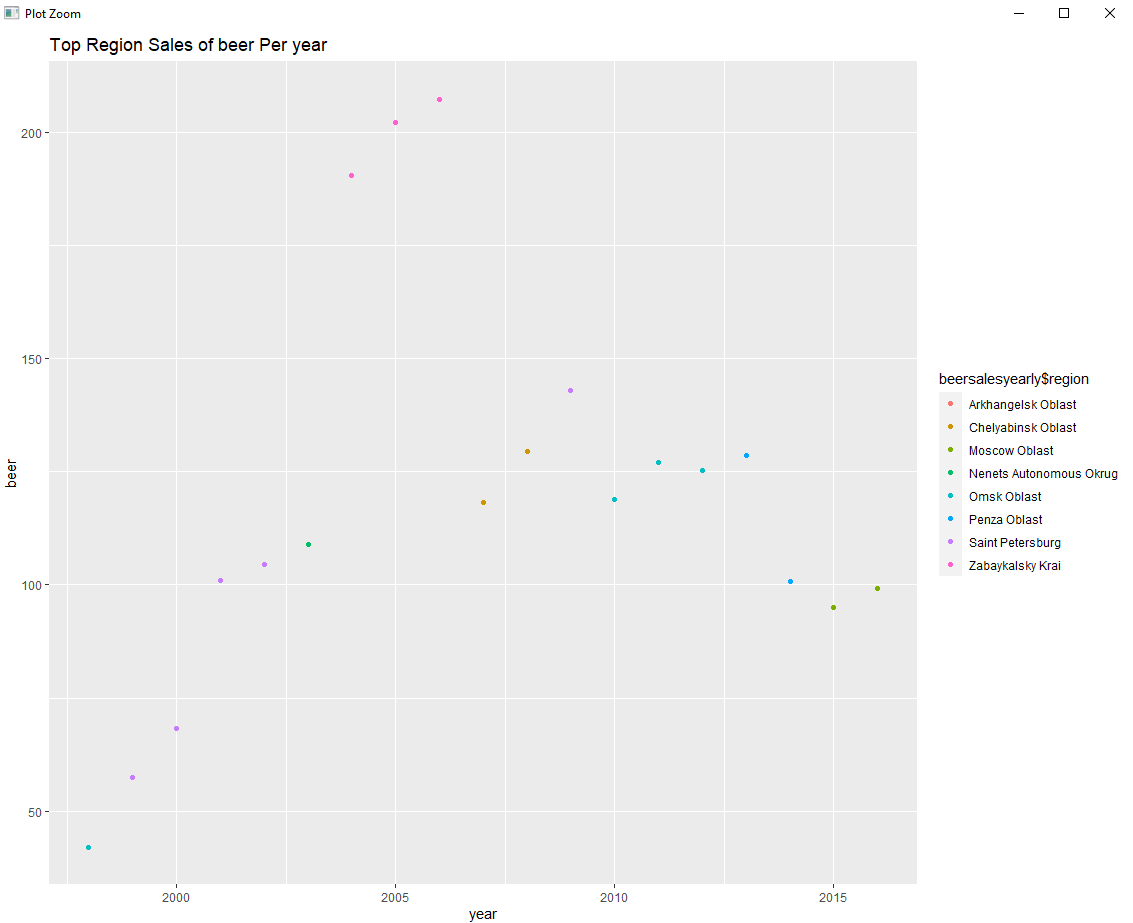
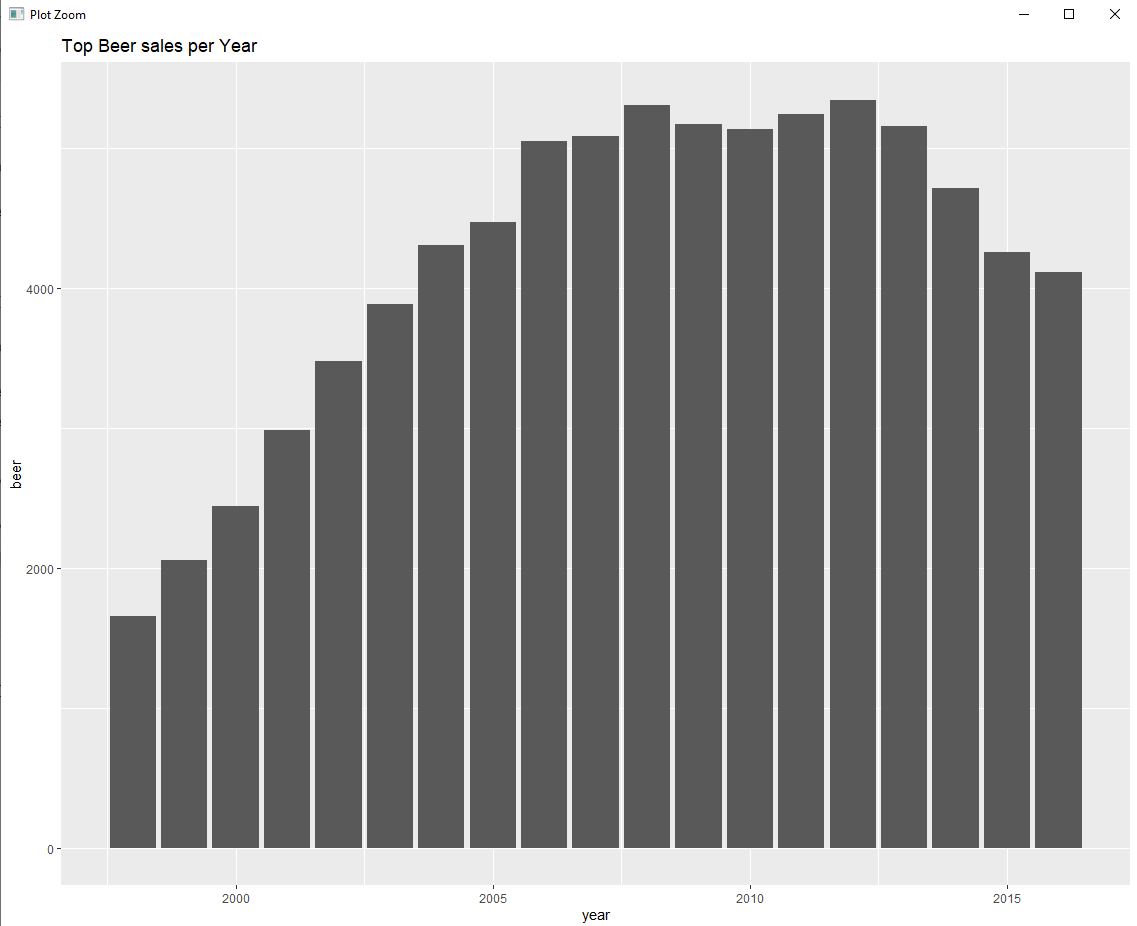


Ilustración 19: Top Beer Region by Year



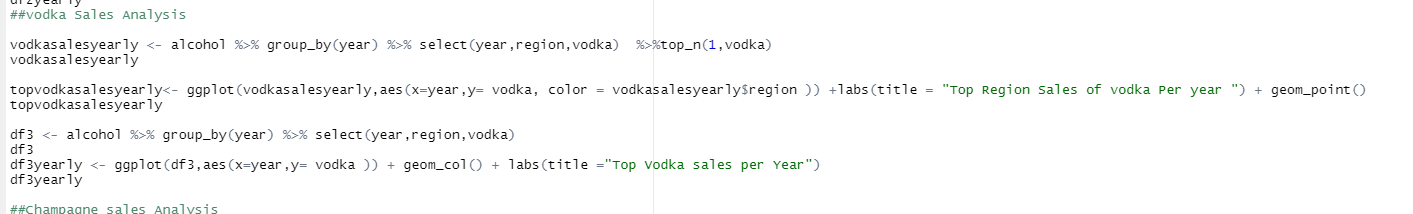


Ilustración 20: Vodka Code

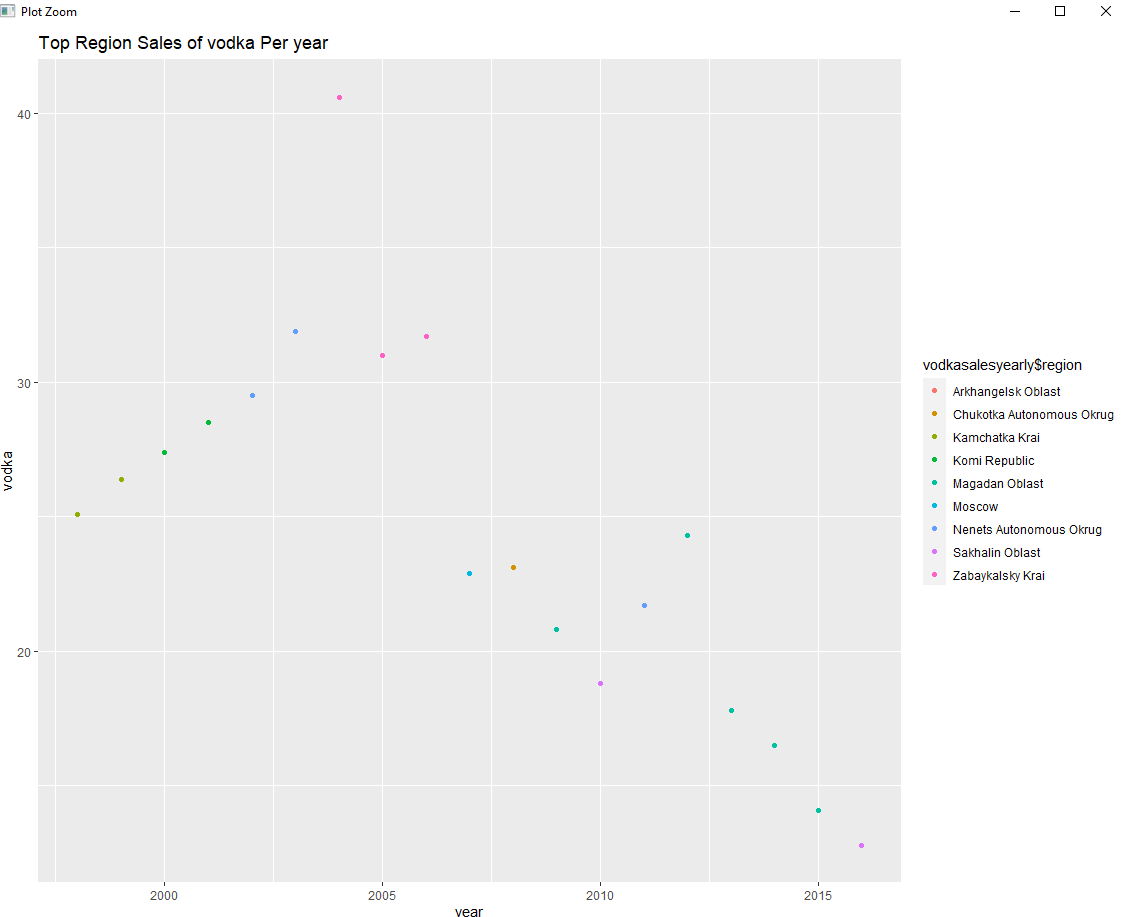


Ilustración 21: Vodka TOP Region Sales by Year

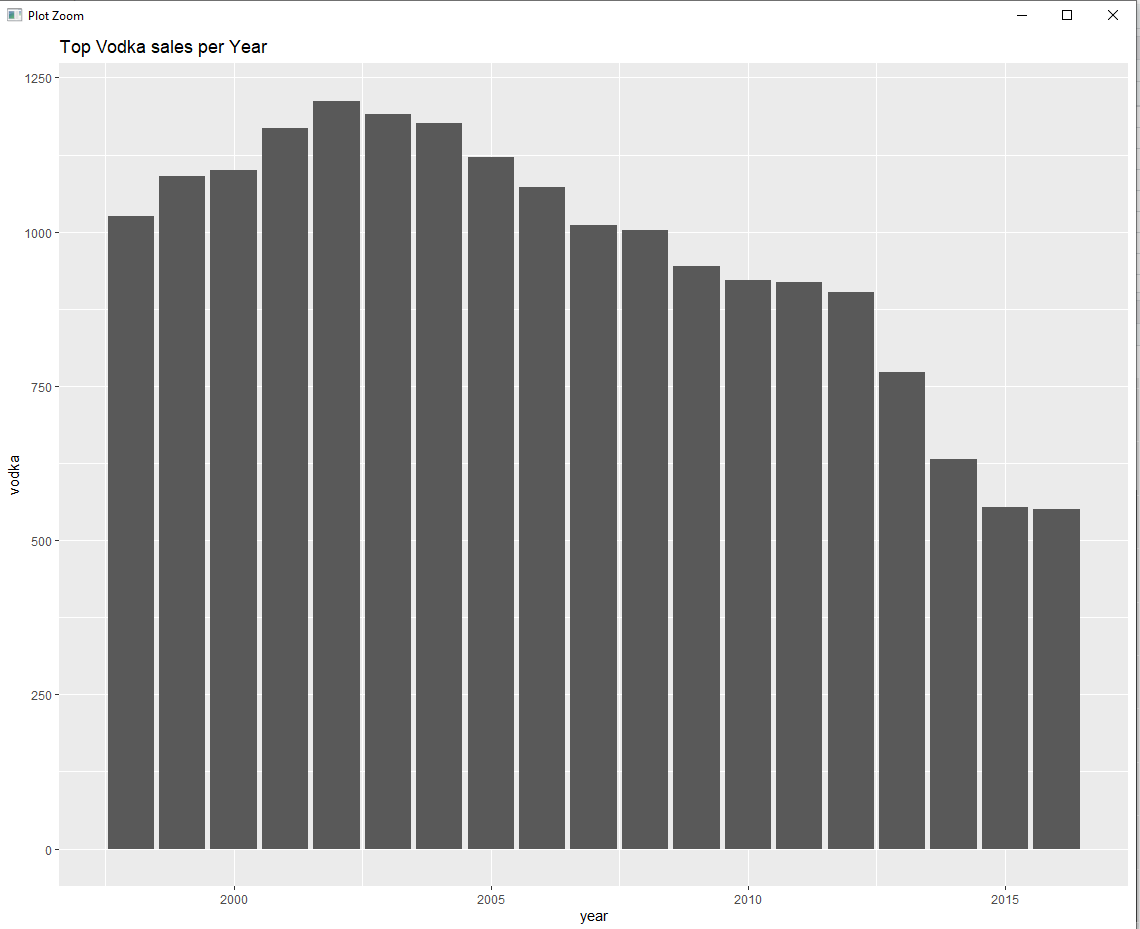


Ilustración 22: Vodka Evolucion Sales by Year

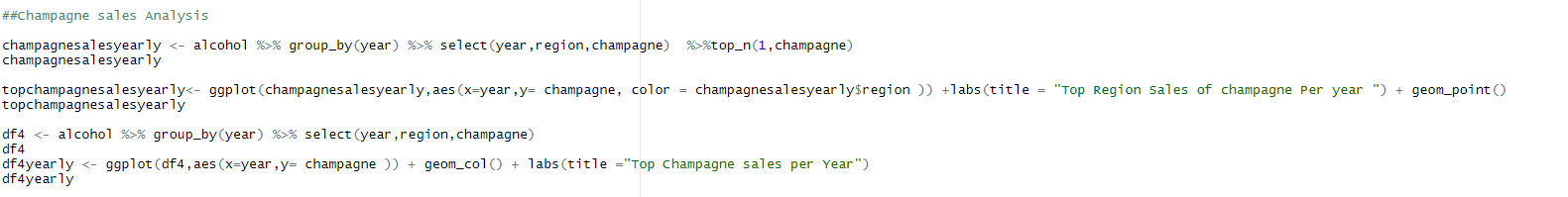


Ilustración 23: Champagne RStudio Code

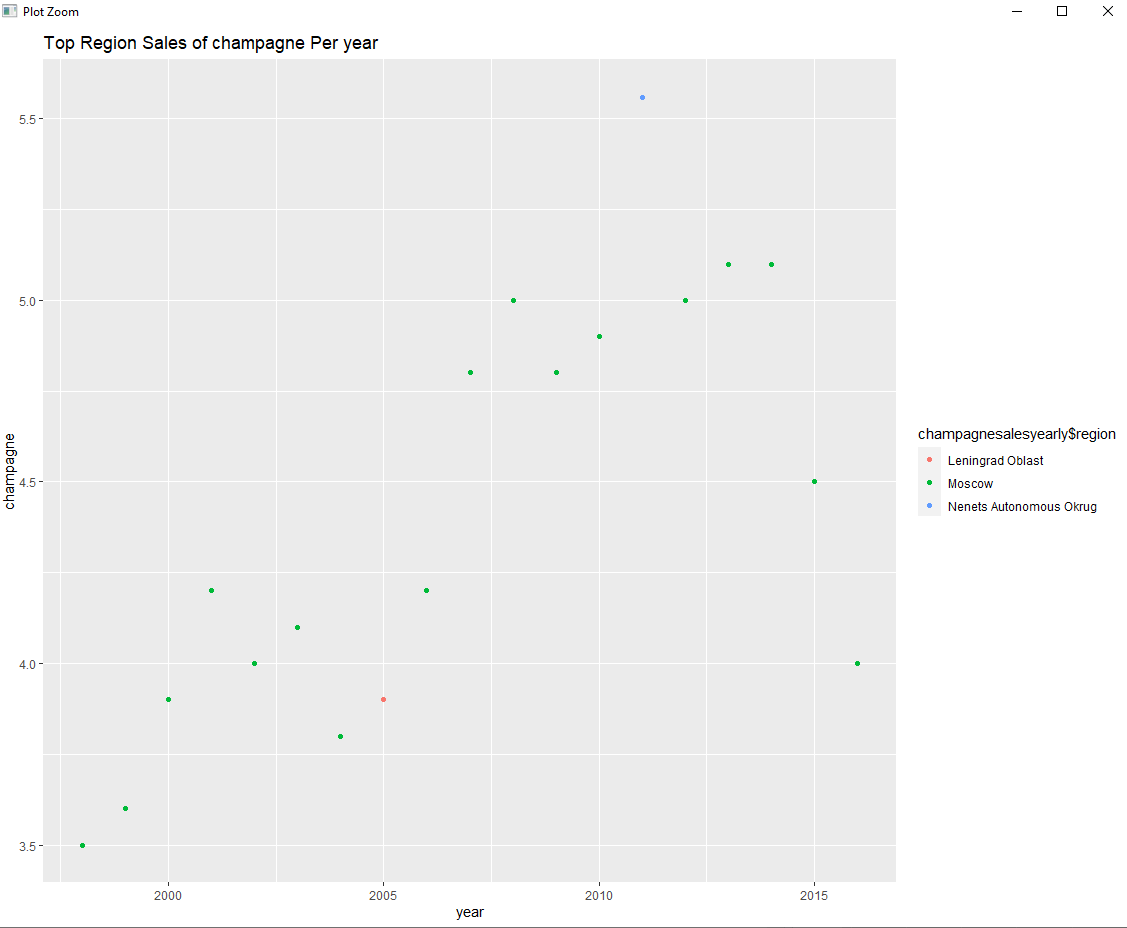


Ilustración 24: TOP Regional Champagne Sales by Year

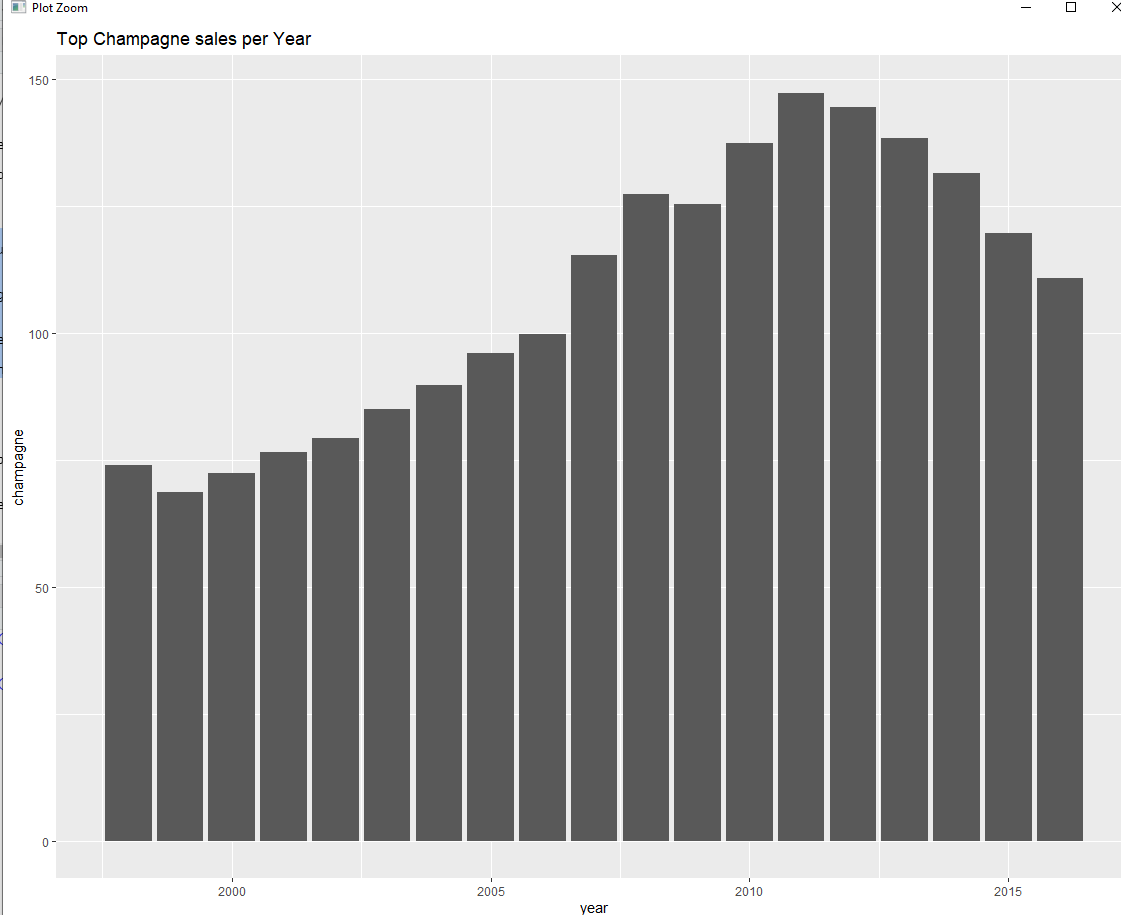


Ilustración 25: Champagne Sales Evolution by Year

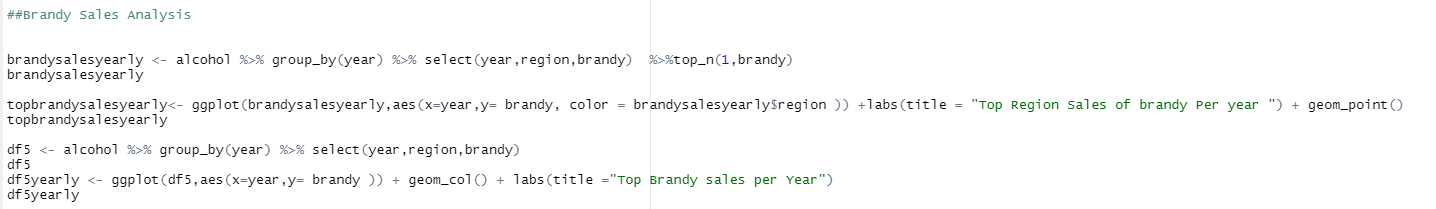


Ilustración 26: Brandy Code View

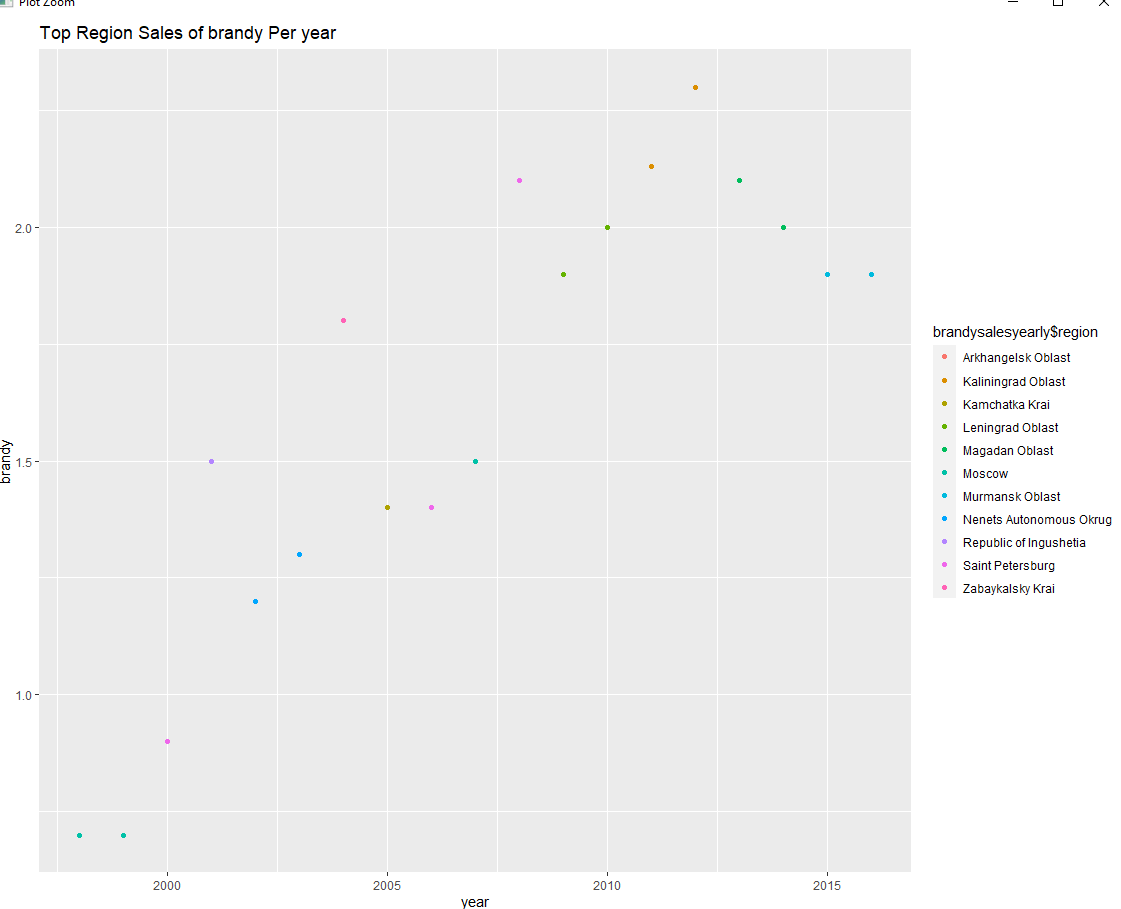


Ilustración 27: Top Brandy Region per Year

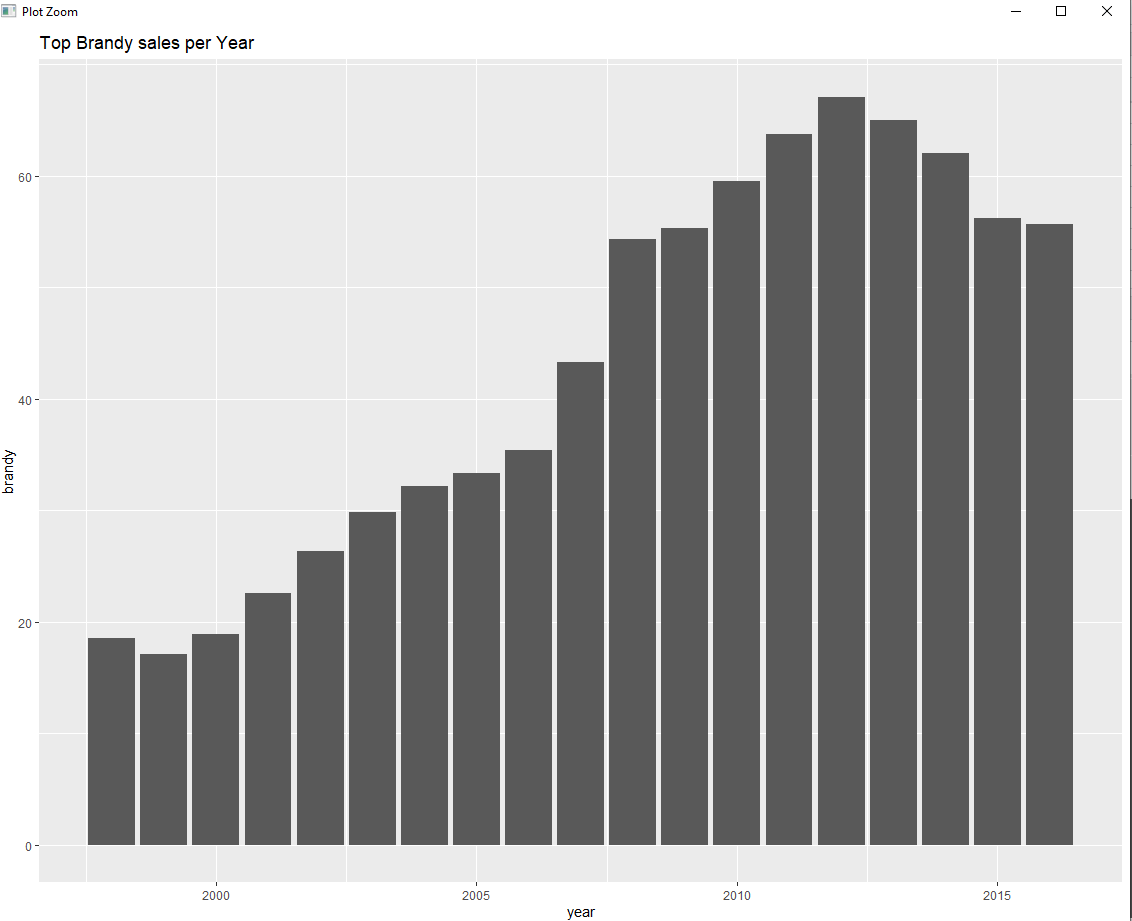


Ilustración 28: Brandy Sales Evolution by Year

A continuacion exploramos la vista boxplot y pairs entender la proporcion de datos outliers y evaluar las posibles correlaciones lado a lado de cada uno de los productos alcolicos. Observandose que el producto beer tiene mas outiers definidos . Por otro lado se observa cierta correlacion positiva entre las venta de Champagne y Brandy.

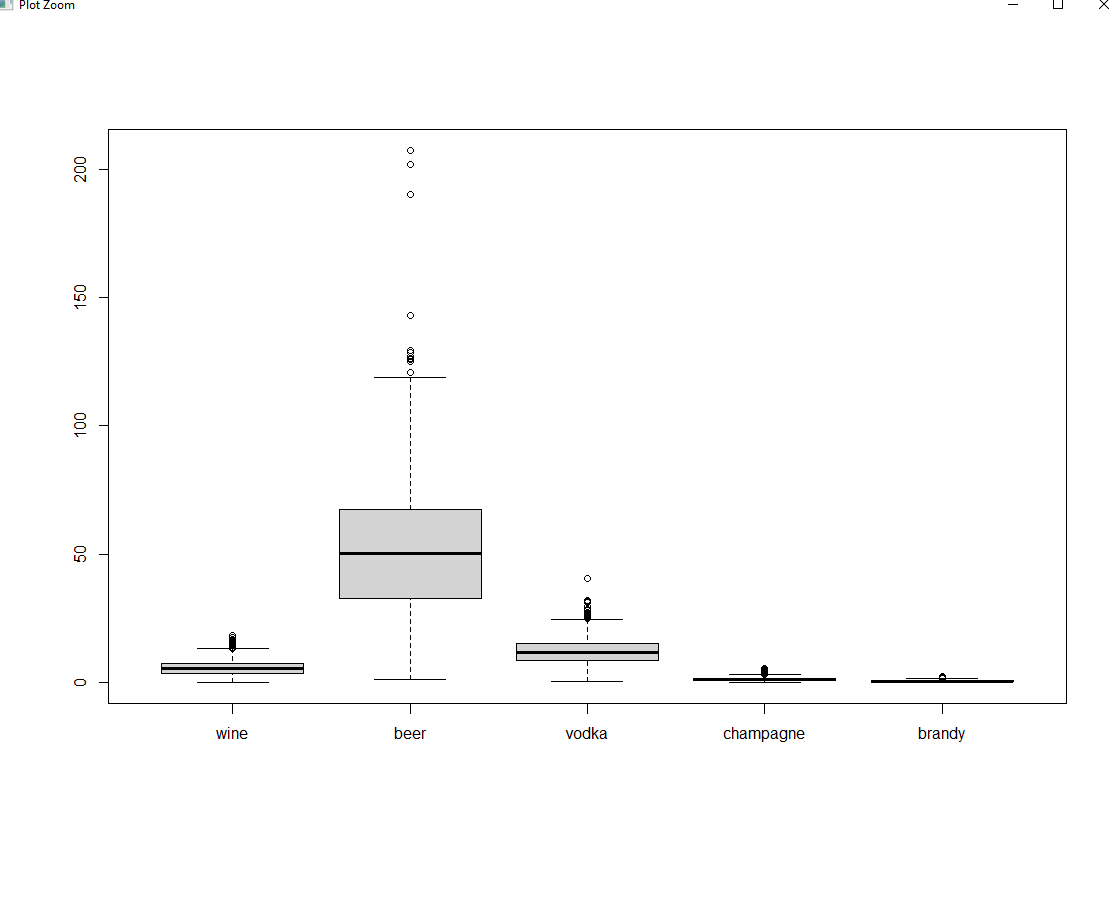


Ilustración 29: Visualizacion Boxplot y Ouliers data

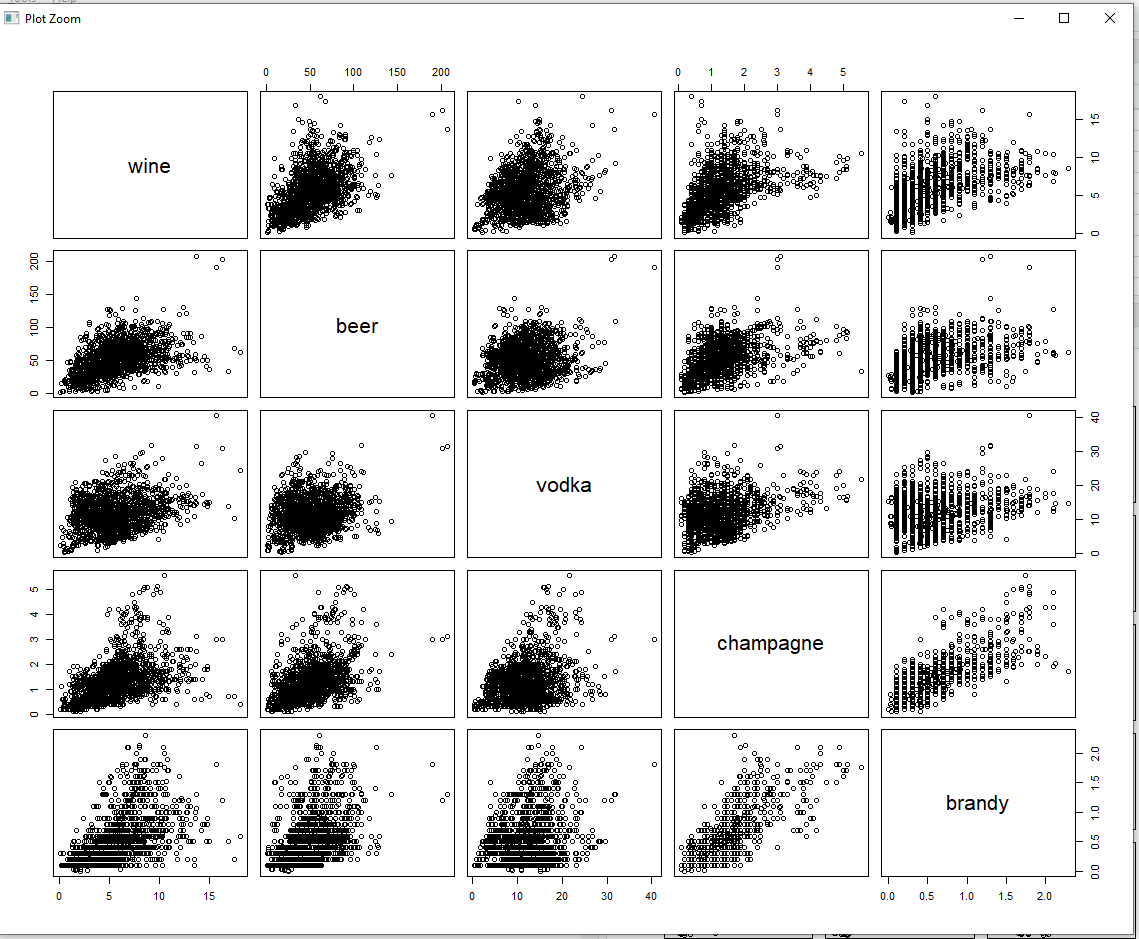
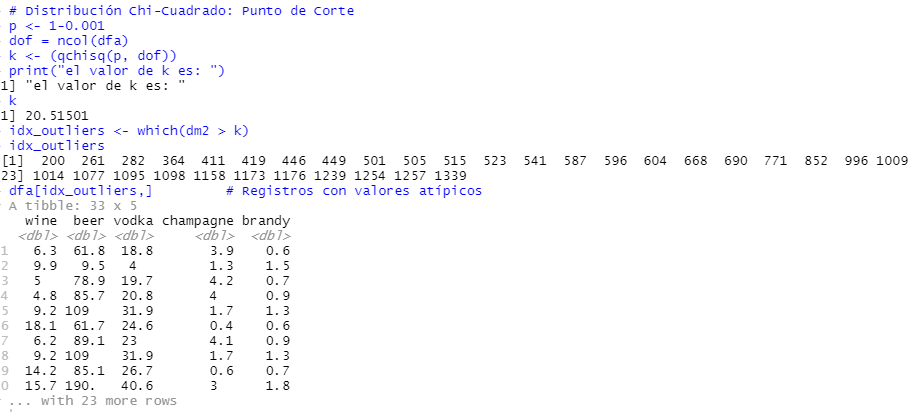


Ilustración 30: Grafico de Correlación de las Ventas de los Productos

Luego se evalúa el puntaje Z para los ouliers además de evaluar los valores atípicos mediante los mecanismos QQ y Chi cuadrado, Mahalanobis donde se obtiene el k= 20.51





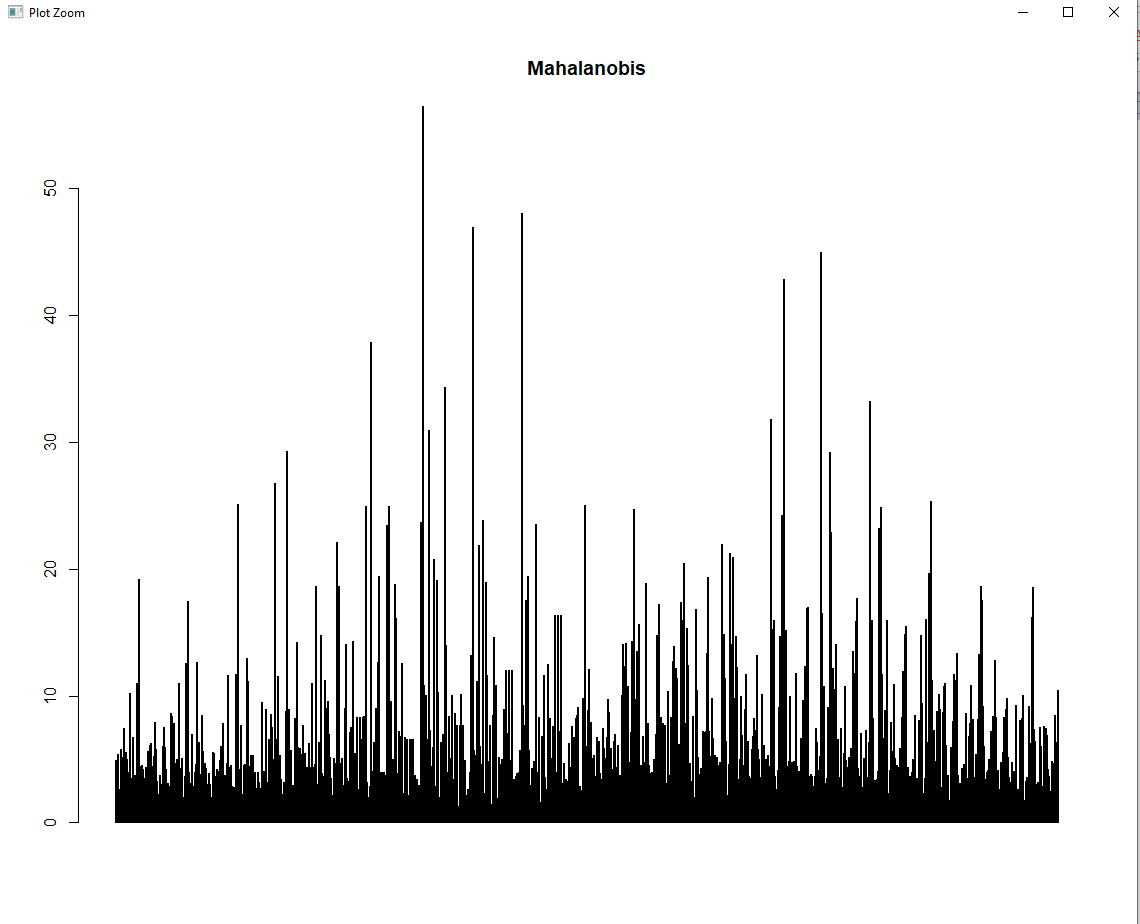
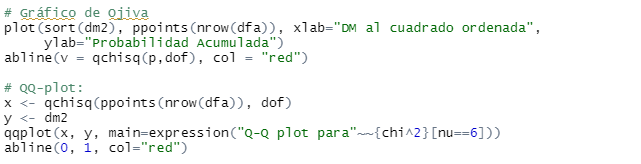


Ilustración 31: Outliers Data

Funcionalmente para visualizar la distribución de los valores atípicos se implementa la codificación para obtener la visualización de ojiva y Q-Q de Chi^2.



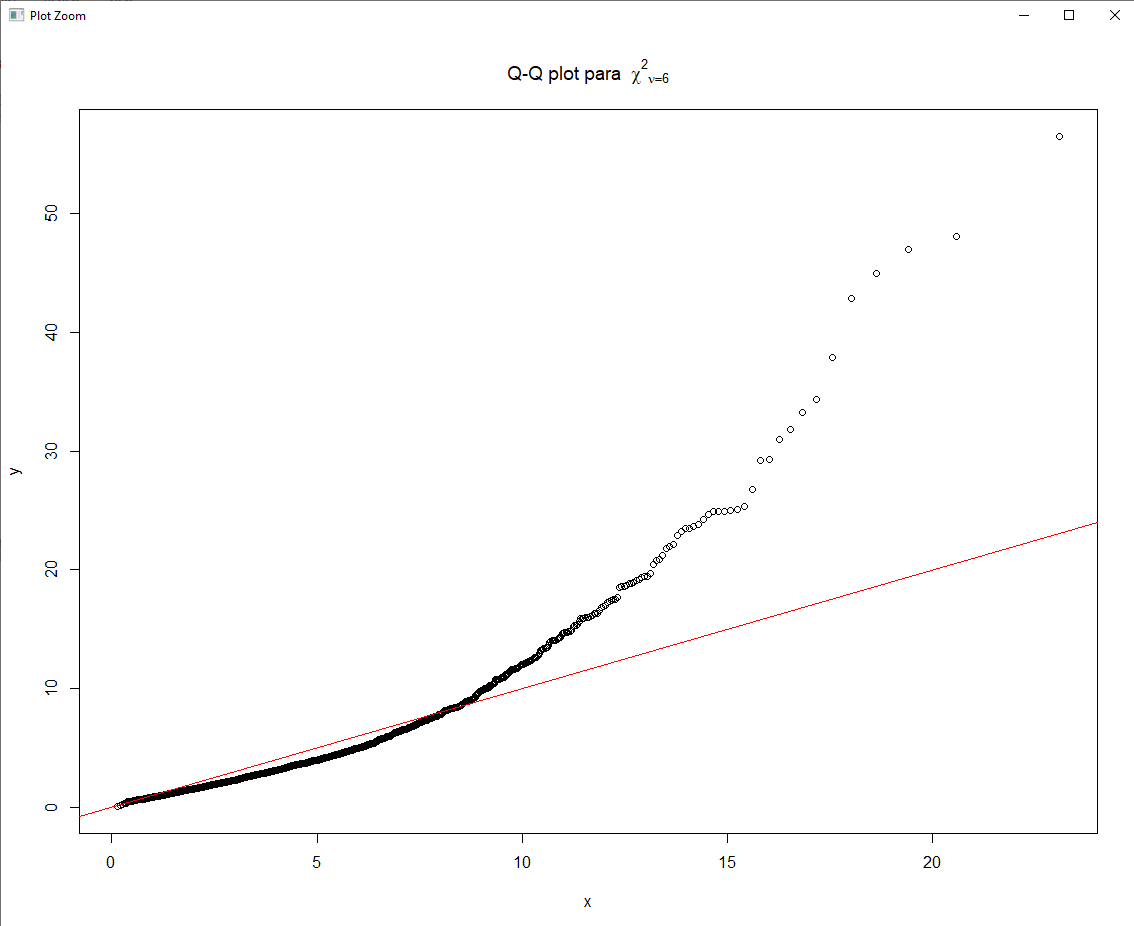


Ilustración 32: Grafico de Q-Q

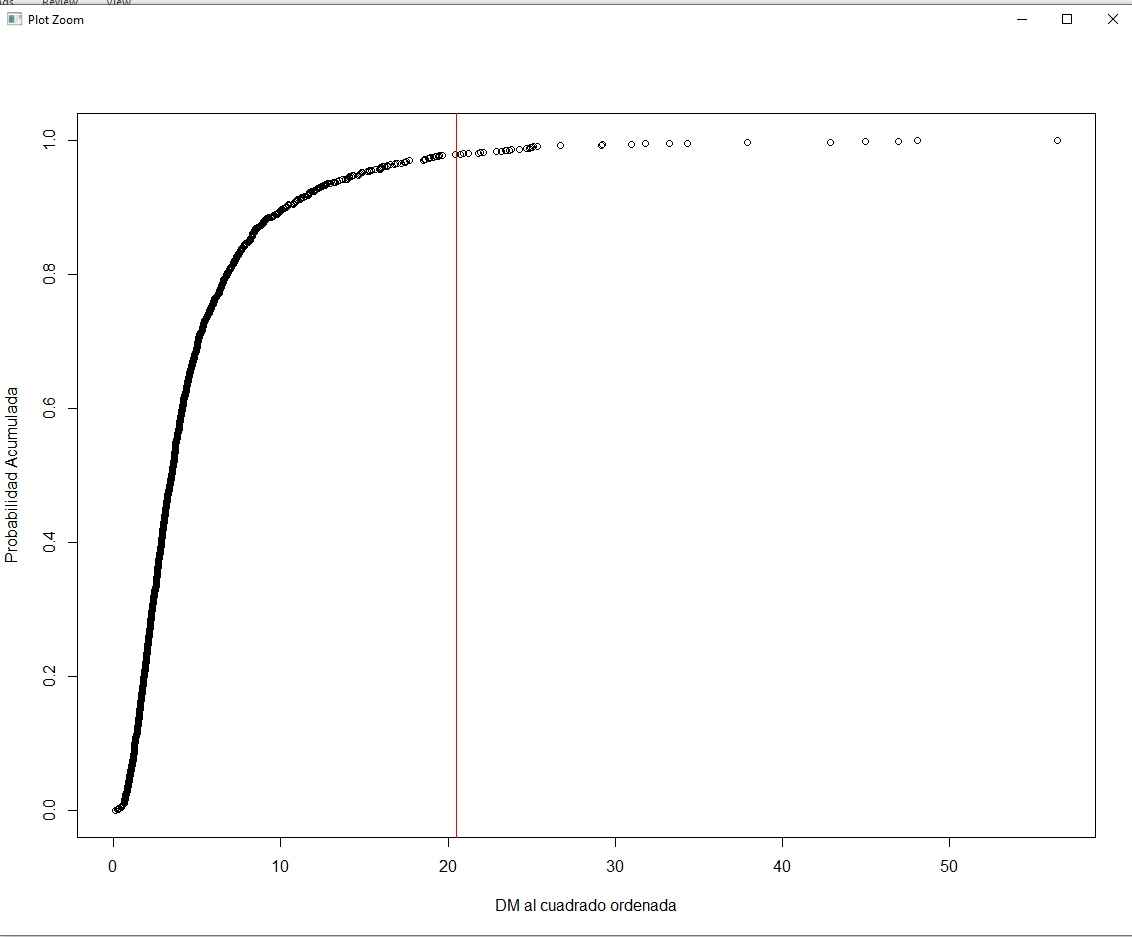
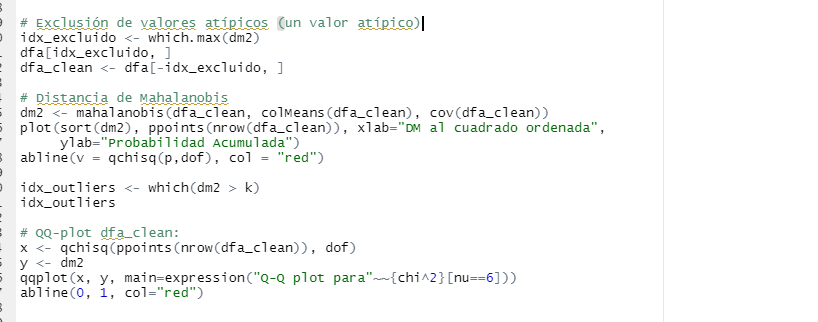
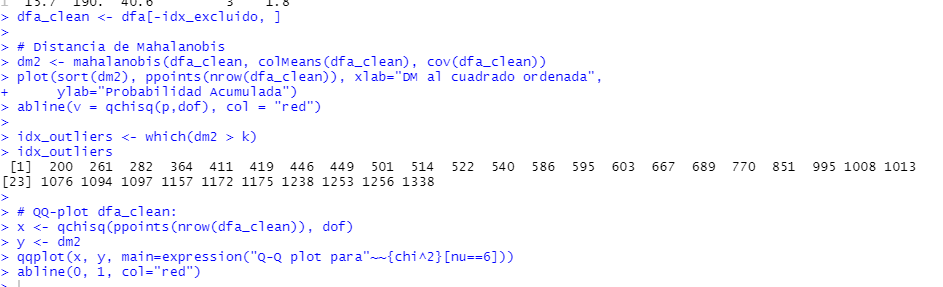


Ilustración 33: Grafico de Ojiva con Outiers

Finalmente, volvemos hacer las mismas visualizaciones descartando los valores atípicos identificados en el paso anterior y volvemos a graficar la Ojiva y Q-Q:





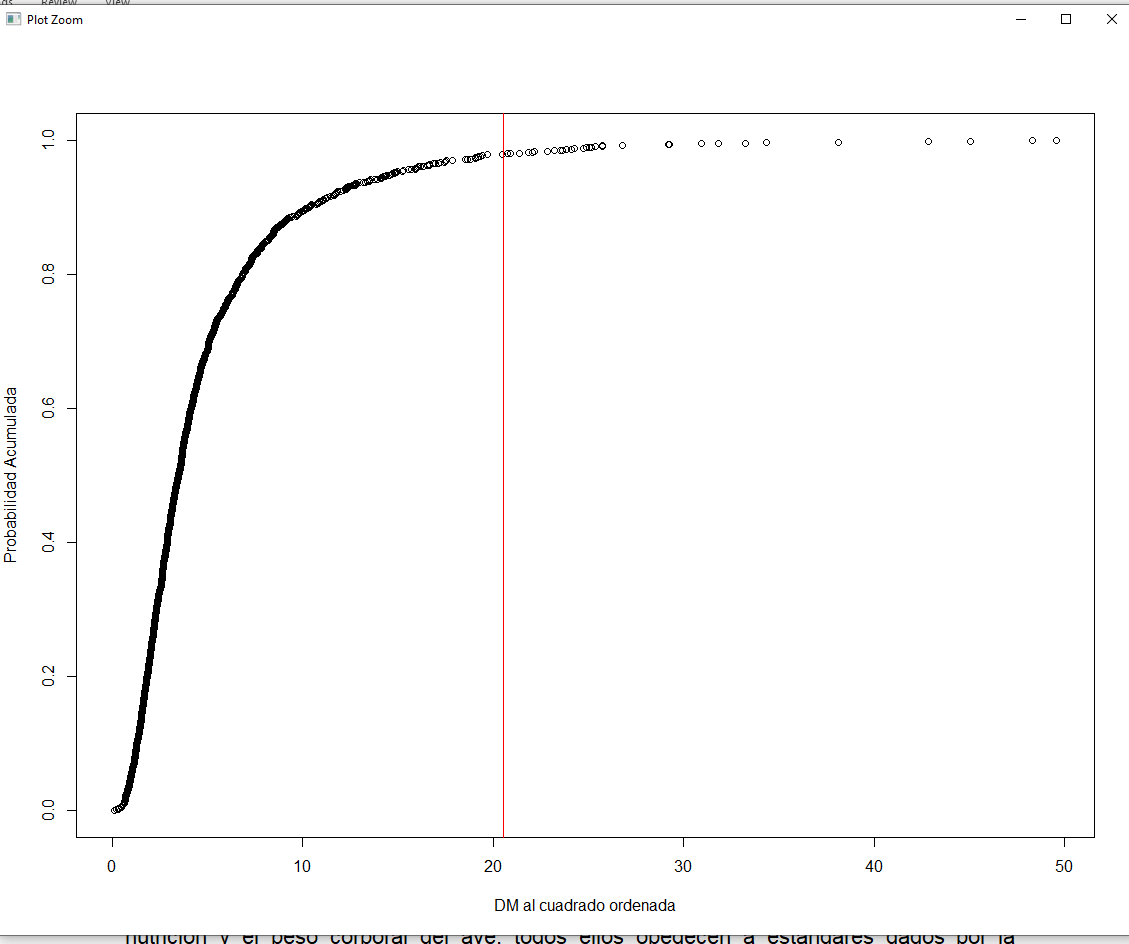


Ilustración 34: Grafico de Ojiva con limpieza de outliers

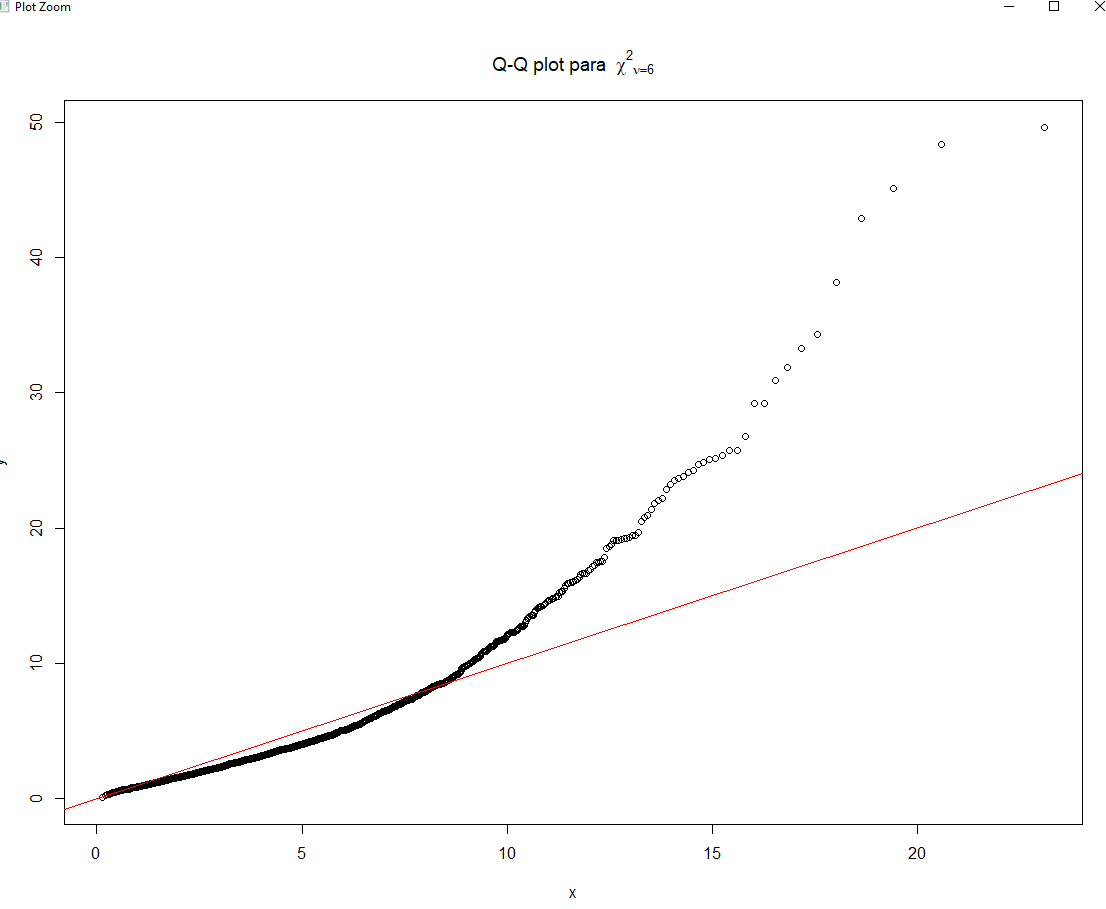


Ilustración 35: Grafico de Q-Q con limpieza de Outliers

## Tercer Caso: Control de Pesos en la Crianza de Aves

El contexto de este caso está en la industria avícola de la empresa peruana Grupo Santa Elena S.A. que tiene como principal actividad la crianza de aves de corral, específicamente pollos engorde. El proceso de crianza de pollos busca lograr el máximo de kilogramos de carne de pollo y para ello debe cuidar sus parámetros productivos como la mortalidad, la nutrición y el peso corporal del ave; todos ellos obedecen a estándares dados por la genética del ave, sistema de crianza y sexo.

El control del peso por lo general se hace en cada semana, pero en la última etapa de la crianza se acorta a dos días. Esto significa que se registran los muestreos por cada galpón de crianza distinguiendo los machos de las hembras a los 7, 14, 21, 28, 35, 28, 40 y 42 días de edad de crianza.

Imagen que contiene Diagrama

Descripción generada automáticamente

Procedimiento del análisis exploratorio a los pesos de control de las aves

**Paso 1:** Se obtuvo los datos de una muestra de 1206 registros en formato CSV con la siguiente estructura:

ID\_GALPON: Identificador del corral de crianza que es una división del galpón que agrupa aves del mismo sexo.

SEXO: Se identifica con “M” a los machos y con “H” a las hembras.

DIA07: Peso del en gramos a los 7 días de crianza.

DIA14: Peso del en gramos a los 14 días de crianza.

DIA21: Peso del en gramos a los 21 días de crianza.

DIA28: Peso del en gramos a los 28 días de crianza.

DIA35: Peso del en gramos a los 35 días de crianza.

DIA38: Peso del en gramos a los 38 días de crianza.

DIA40: Peso del en gramos a los 40 días de crianza.

DIA42: Peso del en gramos a los 42 días de crianza.

La siguiente es una muestra de los datos obtenidos:



**Paso 2**: Mediante codificación Python se logró conectar a la base de datos alojado en el servidor de Atlas Mongo DB.

import pymongo

client = pymongo.MongoClient("mongodb://jgonzalez:rBR3AVVufD5bmQdg@cluster0-shard-00-00.4puwy.mongodb.net:27017,cluster0-shard-00-01.4puwy.mongodb.net:27017,cluster0-shard-00-02.4puwy.mongodb.net:27017/myFirstDatabase?ssl=true&replicaSet=atlas-2q64zi-shard-0&authSource=admin&retryWrites=true&w=majority")

db = client['BD\_GRANJAS']

db.create\_collection('pesos')

**Paso 3:** La carga de datos desde el archivo CSV a la base de datos de MongoDB se ejecutó comandos Python.

import csv

import pandas as pd

data = pd.read\_csv (r'datos.csv',sep=";")

df = pd.DataFrame(data)

lista =[]

row = {"\_id": 1,

'ID\_GALPON':'',

'SEXO':'',

'DIA07': 0,

'DIA14': 0,

'DIA21': 0,

'DIA28': 0,

'DIA35': 0,

'DIA38': 0,

'DIA40': 0,

'DIA42': 0}

n = len(df)

for i in range(n):

row['\_id'] = i

row['ID\_GALPON'] = df.loc[i, "ID\_GALPON"]

row['SEXO'] = df.loc[i, "SEXO"]

row['DIA07'] = df.loc[i, "DIA07"]

row['DIA14'] = df.loc[i, "DIA14"]

row['DIA21'] = df.loc[i, "DIA21"]

row['DIA28'] = df.loc[i, "DIA28"]

row['DIA35'] = df.loc[i, "DIA35"]

row['DIA38'] = df.loc[i, "DIA38"]

row['DIA40'] = df.loc[i, "DIA40"]

row['DIA42'] = df.loc[i, "DIA42"]

db.pesos.insert\_one(row) #Insertando el documento

**Paso 4**: Conectividad y obtención de datos de Atlas MongoDB de la base de datos DB\_GRANJA y la colección pesos

#Recuperando documentos de la colección

mycol = db["pesos"]

#Comprobando

for x in mycol.find().limit(5):

print(x)

**Paso 5**: Lectura y recorrido de los datos obtenidos. El recojo de los datos como Collection se agregó a una lista para luego llevarlo a un objeto DataFrame Pandas

#Recuperando PESOS de las aves Macho

myquery = {"SEXO": "M"}

mydoc = mycol.find(myquery)

data\_list = []

for x in mydoc:

data\_list.append(x)

df\_machos = pd.DataFrame(data\_list)

#Recuperando PESOS de las aves Hembra

myquery = {"SEXO": "H"}

mydoc = mycol.find(myquery)

data\_list = []

for x in mydoc:

data\_list.append(x)

df\_hembras = pd.DataFrame(data\_list)

Paso 6: Explorando datos de los pesos mediante gráficos estadísticos. Se uso la librería matplotlib para plotear las gráficas.

Primero se recuperaron los pesos iniciales de 7 a 28 días.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Las gráficas de frecuencia se sobreponen las distribuciones de los pesos de hembras y machos. Se observa que los pesos de los machos tienen menor variabilidad frente a los pesos de las hembras. Las gráficas evidencian el peso según edad teniendo una media de 181 gramos a los siete días, luego sube a 460 gramos a los 14 días, sube a 881 gramos a los 21 días y llega a los 1394 gramos a los 28 días. A los 28 días ya se empiezan a evaluar a las aves para su saca o entrega como producto terminado como pollo en pie.

Tabla

Descripción generada automáticamente

Gráfico, Histograma

Descripción generada automáticamente

Por separado recuperamos los pesos a partir de los 35 días, es decir en edades disponibles para su comercialización. Se evidencia que la media a los 35 días es de 1990 gramos, pasa a 2210 a los 38 días, llega a los 2339 gramos a los 40 y cierra con 2441 gramos a los 42 días. Son pesos medios entre machos y hembras, lo que sirve para hacer cálculos de rendimiento del proceso por los kilogramos de carne de pollo como resultados.

Texto

Descripción generada automáticamente

Texto

Descripción generada automáticamente

Se observa que los pesos de las aves macho terminan siendo mayores al de las hembras respetando el patrón estándar.

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

En el análisis de los “outliers” por sexo se observa en los pesos de las aves machos que en las primeras semanas hay mucho ruido y esto lo evidenciamos en la siguiente gráfica de cajas con la acentuación de las circunferencias a las 7 y 14 días.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

En el caso de los pesos de las hembras el ruido de los “outliers” se da al inicio y al final de la crianza

Gráfico

Descripción generada automáticamente

## Cuarto Caso Análisis de Outliers de precios en autos usados de la marca BMW

Para el cuarto caso, se analizará la data de venta de carros usados de la marca BMW del año 2020, aplicando las herramientas aprendidas en clase para determinar que atributos son considerados “outliers”. La data contiene información de precio, año de fabricación, modelo, transmisión, millas, tipo de combustible, impuesto, millas por galón y tamaño de motor (en cilindros). Teniendo en cuenta esta información podríamos analizar si los outliers del precio de los autos usados, de la marca BMW, es determinado por el año de fabricación, tipo de transmisión, millas, millas por galón, impuesto y/o tamaño de motor o si existe algún otro tipo de factor que podría estar determinando estos precios.

Model – Modelos de la marca BMW

Year – Año de fabricación

Price – Precio de venta

Tranmission – Tipo de transmisión del vehículo

Mileage – Millas recorridas

fuelType – Tipo de combustible

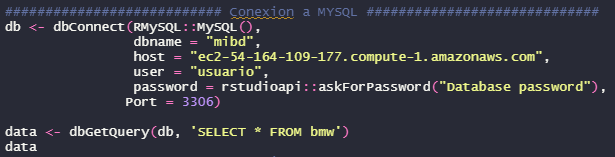
tax – Impuesto anual

mpg – Millas por galón

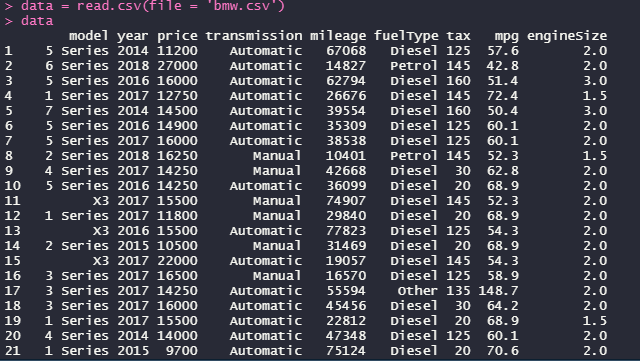
engineSize – Tamaño del motor en centímetros cúbicos

Metodología

Al igual que en los 2 primeros casos, se optó por utilizar el servicio MySQL instalado en una instancia de AWS EC2 con la finalidad de establecer una conexión al entorno de desarrollo y visualización de RStudio, habilitando así el acceso a todos los integrantes del grupo.



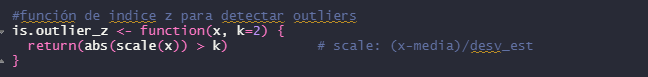
Ejemplo del database bmw.csv



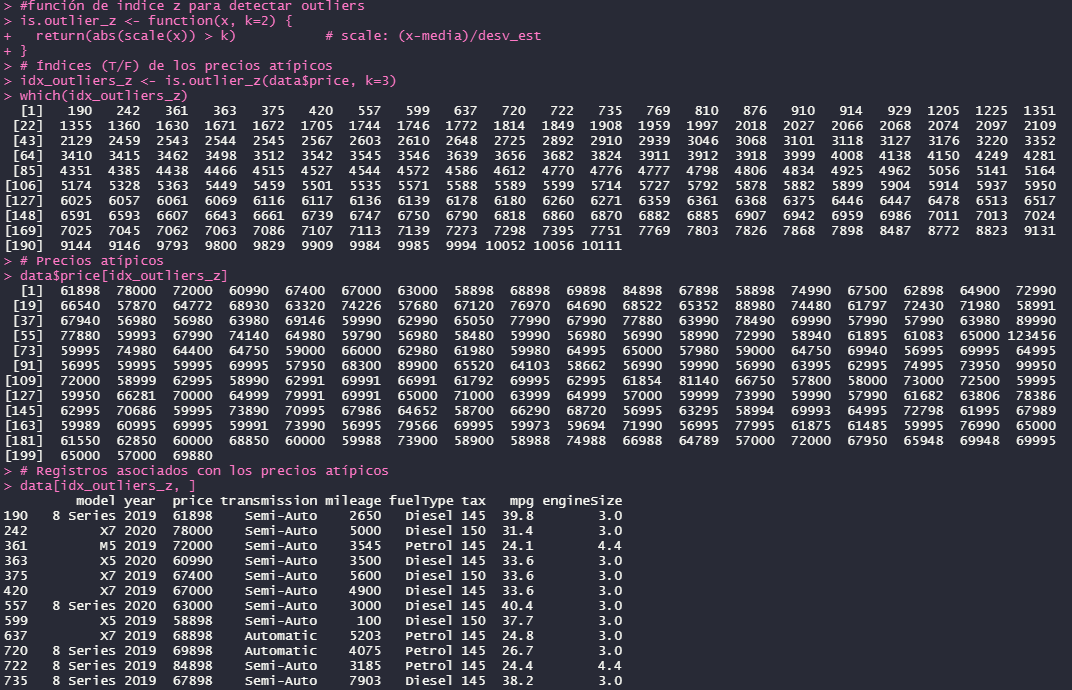
En primer lugar, se verificó que la data no tuviera datos perdidos antes de hacer el análisis de outliers. En este caso, no se encontraron datos perdidos.



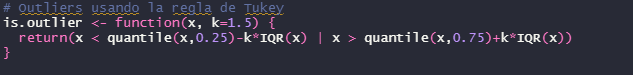
Luego, se realizó el índice Z para detectar outliers de la columna precio.



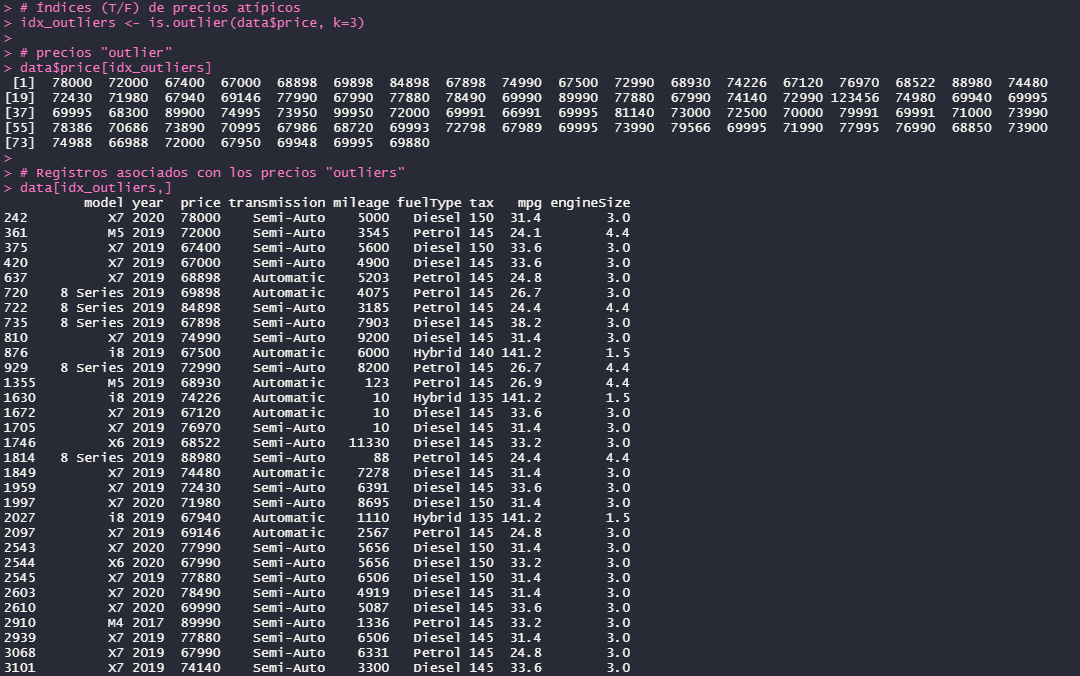
Ejemplo de los precios atípicos obtenidos al realizar el índice Z



Asimismo, se usó la regla de Tukey para verificar con ambos métodos de valores atípicos univariados.



Ejemplo de los precios atípicos obtenidos al realizar la regla de Tukey



Finalmente, se decidió por usar las librerías de ggplot2 y ggpubr para graficar boxplots de los datos outliers para cada caso. Donde se realizaron 6 graficas:

Modelo // Precio

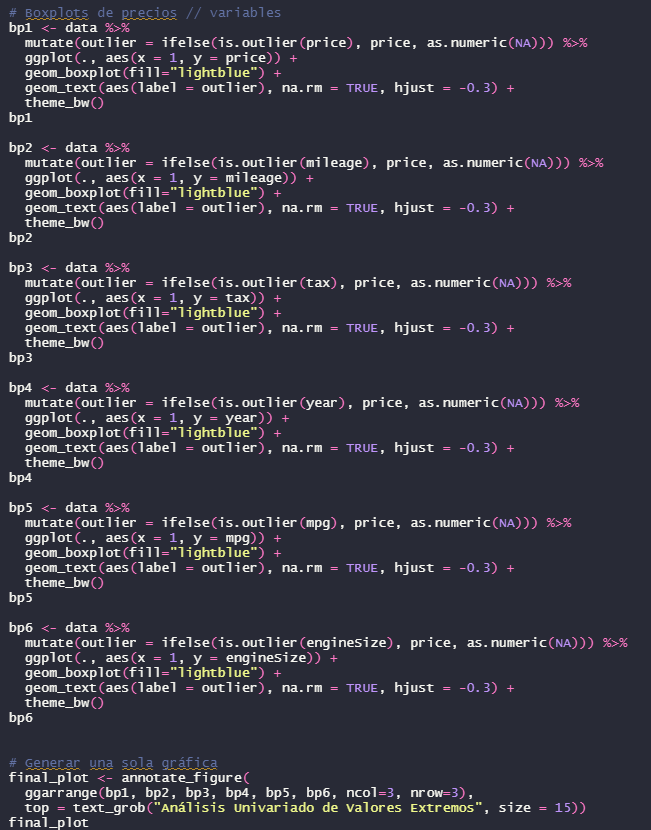
Precio // Millas

Precio // Tax

Precio // Año

Precio // Millas por hora

Precio // Tamaño del motor

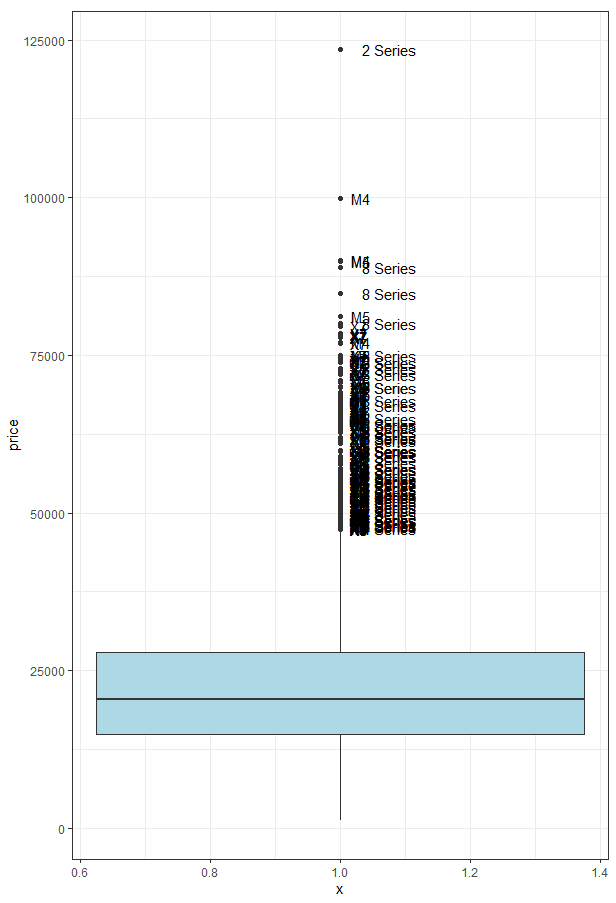


Resultados

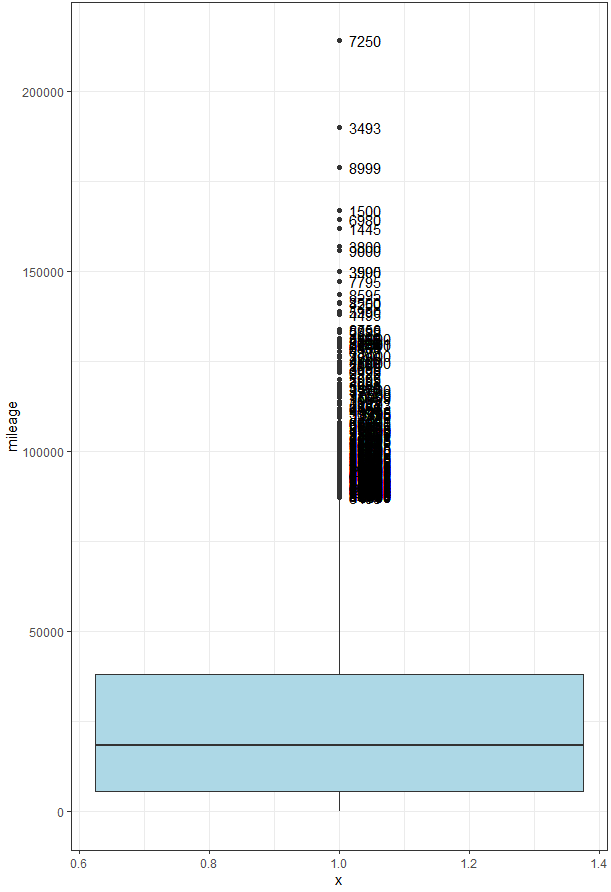
Cómo resultado se esperaba obtener una característica específica que determine el precio por encima del promedio en la venta de autos usados de la marca BMW. Y si bien se puede concluir que el precio del auto será mayor mientras más reciente sea el año de fabricación, menos millas tenga el vehículo y más grande sea el motor. También existen casos en donde aparenta haber otro atributo que no se encuentra en la base de datos que este influenciando el precio. Por otro lado, se puede pensar que hubo errores al ingresar la data, como por ejemplo con el auto que fue vendido por el precio más alto:



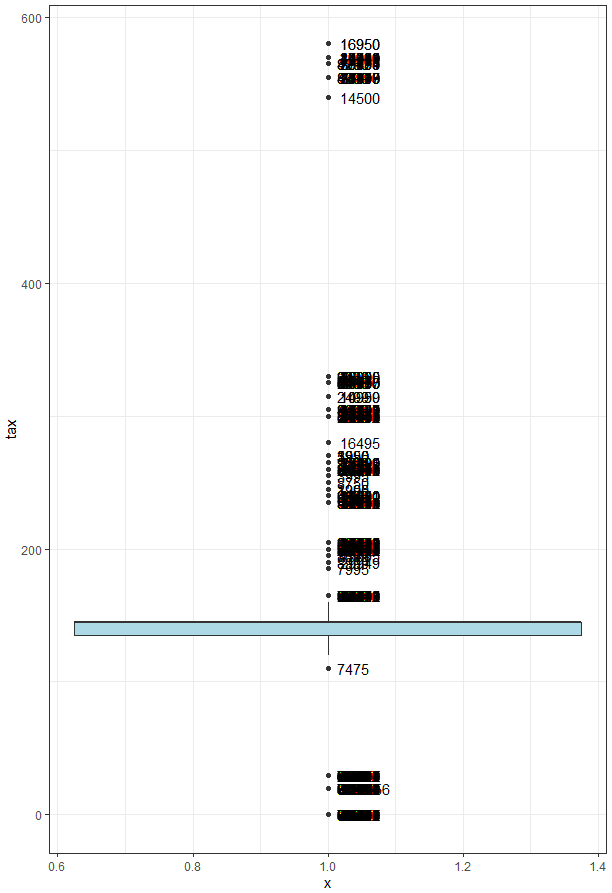
Precio = $123456



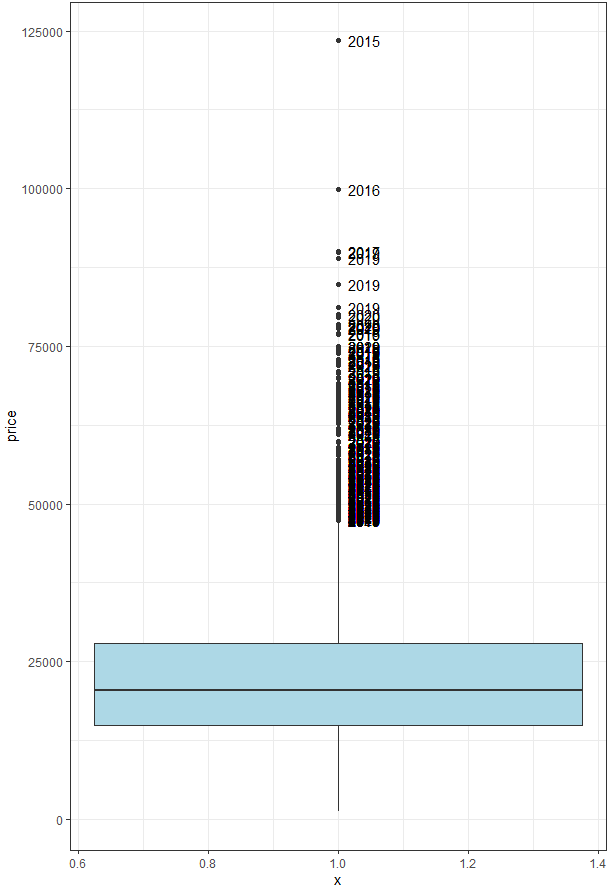
Boxplot 1 precios atípicos sobre modelo de BMW



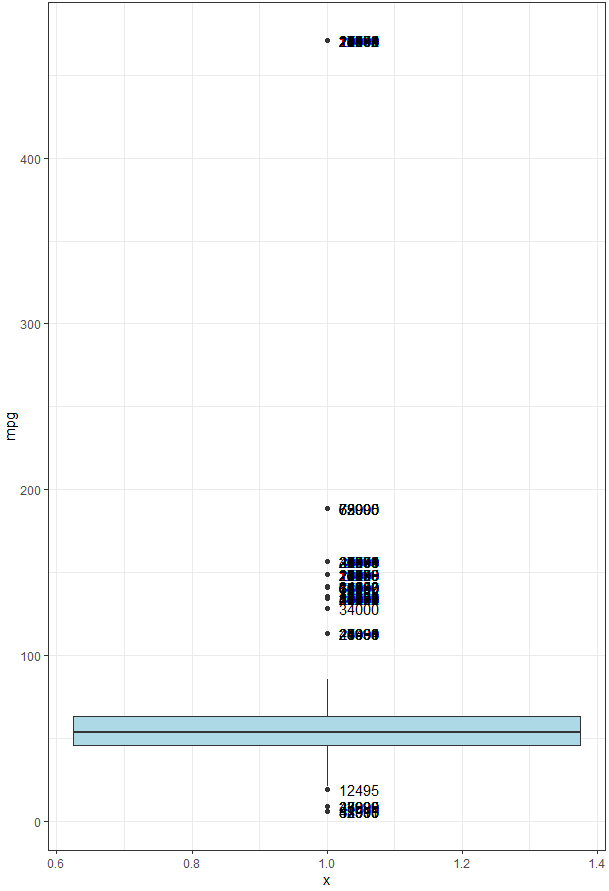
Boxplot 2 precios atípicos sobre millas acumuladas del vehículo



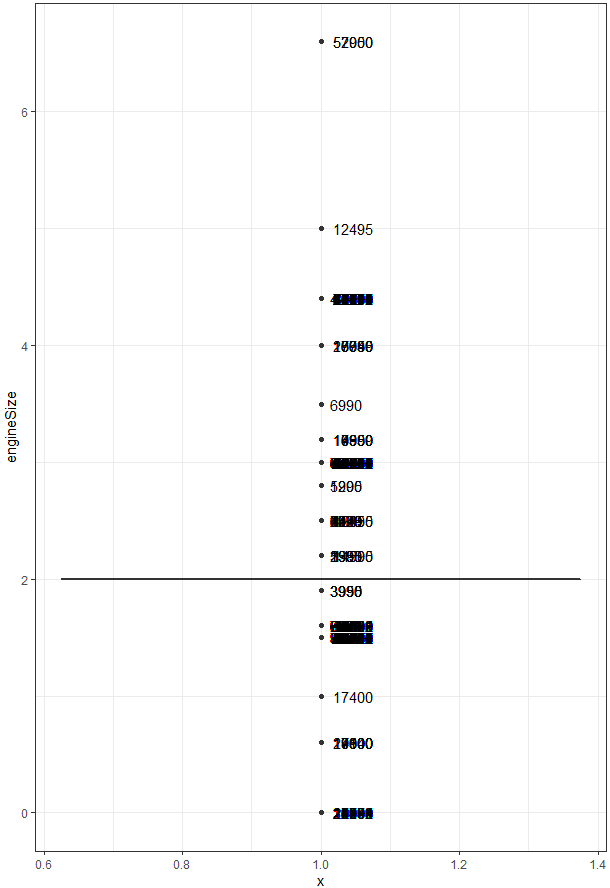
Boxplot 3 precios atípicos sobre impuesto anual



Boxplot 4 precios atípicos sobre año de fabricación



Boxplot 5 precios atípicos sobre millas por galón



Boxplot 6 precios atípicos sobre tamaño de motor

# Conclusiones y Recomendaciones

Primer caso: Se replica la resolución de la hipótesis planteada por el dr.Semmelweis que sostuvo que el lavado de manos previa antes de una nueva atención de parteras reducía la mortandad de los niños en el SXIX, pues ser redujo la proporción de mortandad de 10.5% a 2.1%.

Segundo Caso: Se utilizan diversos esquemas de visualización interactiva, gestión de datos perdidos, análisis outsiders, para ofrecer al área de marketing las top 10 regiones idóneas para replicar las estrategias de mercadeo que fueron exitosas en Saint Petersburg.

Tercer Caso: Se realiza un análisis exploratorio de los pesos de las aves vivas aplicando graficas de distribución, bloxplot, etc. para detectar los outliers.

Cuarto Caso: En síntesis, se puede confirmar la hipótesis de que los atributos de año, millas y tamaño de motor son los que más interfieren en un precio por encima del promedio de los vehículos. Igualmente, se recomienda hacer otro tipo de análisis, cómo por ejemplo clustering, para hallar una predicción de precio más exacta con la data. Otra observación o recomendación que se encontró, es que dentro del atributo de “fuelType” se encontraron 36 vehículos con la variable “Other”, esto bien podría ser datos perdidos u otras fuentes de combustible, como por ejemplo gnv o glp, pero que no están explicitadas en la base de datos.

En relación a la tecnología se creyó convenientes utilizar los servicios de bigdata basados en nube como Atlas Mondo DB y MySQL/EC2 on AWS ya que ofrecen agilidad, rápida escalabilidad y flexibilidad para implementar nuevos casos de uso si la necesidad de involucrarse en proyectos complejos de implementación de centro de datos onpremises

La experiencia y usabilidad de Atlas Mondo DB y el entorno Google Collab Python ha sido satisfactoria a pesar de utilizar el free tier. Al mismo tiempo los servicios MySQL desplegados en AWS EC2 no obstante el DNS que identifica públicamente la EC2 Free tier cambia dinámicamente por la propia naturaleza del servicio. Por esta razón, se recomienda adquirir el servicio de Route53 para poder registrar un dominio estático y por consiguiente no se requeriría actualizar las conexiones de Rstudio y Powershell.

Las librerías gráficas estadísticas tanto en Rstudio y en Google Collab Python son de fácil uso, despliegue y ayudan a la rápida exploración y visualización de datos

# Bibliografía

* Hoja resumen de Shiny: <https://shiny.rstudio.com/articles/cheatsheet.html>
* Funciones Rstudio-Shiny: <https://shiny.rstudio.com/reference/shiny/1.6.0/>
* Irizarry, R. A. (2019). Introduction to data science: Data analysis and
* prediction algorithms with R. CRC Press: <https://rafalab.github.io/dsbook>
* AWS EC2 Service: <https://aws.amazon.com/es/ec2/?trk=58ace84c-cd27-448f-9f64-ec1187db737b&sc_channel=ps&sc_campaign=acquisition&sc_medium=ACQ-P|PS-GO|Brand|Desktop|SU|Compute|EC2|LATAMO|ES|Text&s_kwcid=AL!4422!3!590500029733!e!!g!!aws%20ec2&ef_id=Cj0KCQjwma6TBhDIARIsAOKuANxKSCLlz9mI94wIRc7ddYjXGrTZJ-6GJAYryvit2tA0g0TUyCmQKp8aAvhpEALw_wcB:G:s&s_kwcid=AL!4422!3!590500029733!e!!g!!aws%20ec2>
* Documentación Mongo Atlas Service : <https://www.mongodb.com/atlas>

# Anexos

* GitRepository: <https://github.com/u712596/TrabajoFinal>
* CollabRepository Tercer Caso: <https://colab.research.google.com/drive/1-XuGPiT2HIv_KU48hFXHGWNnO6S6Mv14?usp=sharing>

Shiny links Section:

<https://kylixjac.shinyapps.io/clinica/>