## Business Understanding

Diariamente en una aerolínea se toman múltiples decisiones. Estas decisiones afectan de manera negativa o positiva la operación y/o el desempeño de la aerolínea tanto a corto como a largo plazo. Una de las decisiones más importantes que se toman es el precio disponible de un tiquete, el cual se define para cada uno de los Origen-Destino, con diferentes restricciones para los diferentes clientes.

Las decisiones de los precios en una aerolínea afectan directamente los ingresos de la compañía, es por esto que el pricing (precio disponible) se debe considerar como un arma estratégica, la cual permite obtener ventajas competitivas en el mercado (Christ, 2011). Cada una de las decisiones que se toman en un momento particular afecta directamente el futuro de la demanda, por lo que es necesario modificar dinámicamente el precio durante el tiempo para poder maximizar los ingresos. Las aerolíneas tienen el gran reto de fijar el precio óptimo en cada periodo de tiempo de tal manera que se aproveche la disposición de pago de cada uno de los clientes y se vendan cada uno de los tiquetes con la mejor tarifa posible.

Tradicionalmente las aerolíneas Legacy manejan sistemas de Revenue Management que son basados en decisiones de asignación de capacidad como la herramienta primaria táctica para manejar la demanda, es decir, en sistemas tipo quantity-based. Esta estrategia de venta se fundamenta en una demanda Yieldable, lo que significa que los pasajeros se guían por un producto específico o clase de reserva. Pero con la llegada de las aerolíneas Low Cost (LCC) y la masificación del internet, se ha incrementado la transparencia de los precios y los vuelos, causando que los consumidores puedan comparar los precios de la diferente oferta existente y así armar sus planes de viajes tomando ventaja de las tarifas más bajas (Mumbower, Garrow, & Higgins, 2014). Es por esto que la estructura de la demanda ha cambiado y ha ido evolucionando hacia una demanda Priceable, es decir que el cliente busca el menor precio para su viaje.

Hoy en día, Avianca Holdings opera los mercados de Norteamérica, Centroamérica, Sudamérica, Europa, Domestico Colombia, Domestico Perú y Domestico Colombia, cubriendo aproximadamente 800 Orígenes-Destino significativos con 292 rutas. Estos mercados han fluctuado bastante teniendo en cuenta la llegada de aerolíneas Low Cost como VivaColombia que se estableció inicialmente en el mercado Doméstico Colombia desde el año 2009 y desde entonces ha ido creciendo hasta llegar actualmente a operar algunas rutas internacionales como BOGLIM, BOGPTY, BOGMIA, BOGUIO. Interjet y Jetblue también han llegado a estos mercados con rutas como FLLCTG y BOGMEX. Por otra parte, existe una gran competencia con aerolíneas como COPA y LATAM en los mercados de América, IBERIA en el mercado de Europa, y la llegada de nueva competencia como KLM y AirEuropa en el mercado de Europa generando la disminución de los precios de los tiquetes.

De acuerdo a (Talluri & Van Ryzin, 2005) la mejor estrategia para poder analizar la demanda Priceable es Dynamic Pricing. Esta estrategia consiste en ajustar el precio de los tiquetes de manera oportuna y respondiendo a cambios en la demanda y/o acciones del competidor con el propósito de proveer el producto correcto, al cliente correcto, en el momento correcto, al precio adecuado. (Talluri & Van Ryzin, 2005) argumentan que un modelo quantity-based opera bajo el supuesto de controlar la cantidad vendida a los diferentes segmentos de clientes, lo que puede llevar a menos ventas por la limitación de la oferta. Por otro lado, si se tiene flexibilidad con los precios se puede controlar la demanda de una manera más rentable, especialmente en ambientes dinámicos donde la cantidad demanda puede cambiar teniendo un precio constante (Balvers & Cosimano, 1990).

En la actualidad, Avianca Holdings cuenta con un sistema de Revenue Management de tipo quantity-based que determina la disponibilidad de cada O&D y por ende el precio de cada uno de los tiquetes. Los especialistas del área de Demanda que, basándose especialmente en conocimiento previo sobre el mercado, realizan influencias al sistema para modificar la estimación de acuerdo a movimientos del mercado que el sistema no puede capturar o que se demora en tenerlos en cuenta para incorporarlos en la optimización. Actualmente, los especialistas no cuentan con una herramienta de Dynamic Pricing y toman sus decisiones basados en los resultados de años pasados. Sin embargo, estos resultados no siempre son los más apropiados para tomar decisiones ya que en primer lugar, las condiciones del mercado varían de año a año, y en segundo lugar, no se llega a saber a ciencia cierta si la estrategia de los años pasados fue la correcta.

Teniendo en cuenta lo descrito anteriormente, es importante poder contar con una herramienta de Dynamic Pricing con la que los especialistas puedan tomar decisiones acertadas y fundamentadas en hechos y datos analíticos, ya que no se cuenta con alguna herramienta que dé soporte en la toma de decisiones a la hora de realizar influencias. Es importante aclarar que esta herramienta será de soporte y no se piensa como un reemplazo de los sistemas actuales, por el contrario, la idea es que se logren combinar las dos estrategias. El área de Demanda es la encargada de llevar esto a cabo.

### Business Objectives

Por lo tanto, la presente investigación tiene como **objetivo** construir un modelo de Dynamic Pricing que tenga como componentes principales la estimación correcta de la demanda dado un precio y un modelo de programación dinámica. Este modelo tiene el propósito de desarrollar una herramienta que represente como los clientes toman decisiones en el ambiente actual, determinando que precios ofrecer en el mercado para maximizar los ingresos.

También se dará respuesta a las siguientes preguntas:

¿Cuáles son las variables más importantes que influyen en la demanda que se deben tener en cuenta para análisis y estrategias futuras de venta?

¿Cuál es el precio disponible óptimo al que debe venderse un tiquete según su anticipación de compra?

### Data Mining Goals

Para lograr el objetivo principal, es necesario:

* Estimar la curva de la demanda dado un nivel de agregación, que tenga un de mínimo 65%.
* Formular un Dynamic Programming determinístico que se alimente del output de la estimación de la curva de demanda y otorgue el precio óptimo al que debe venderse

### Project Plan

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Fase** | **Actividad** | **Fecha Inicio** | | **Fecha Fin** | **Entregable** |
| **1** | Descripción del problema | | 01/09/2016 | 09/09/2016 | Documento |
| Revisión Bibliográfica | | 12/09/2016 | 16/09/2016 | Documento |
| Entendimiento datos Curva de Demanda | | 19/09/2016 | 23/09/2016 | Datos/Documento |
| Preparación de datos Curva de Demanda | | 26/09/2016 | 05/10/2016 | Datos/Documento |
| Modelaje de Curva de Demanda | | 06/10/2016 | 31/10/2016 | Modelo/Documento |
| Evaluación de Resultados | | 06/10/2016 | 31/10/2016 | Resultados/Documento |
| **2** | Formulación del Modelo de Dynamic Programming | | 01/11/2016 | 07/11/2016 | Documento |
| Preparación Datos | | 08/11/2016 | 30/11/2016 | Datos/Documento |
| Modelaje Optimización | | 03/12/2016 | 31/12/2016 | Modelo/Documento |
| Evaluación y Entrega Final | | 03/12/2016 | 31/12/2016 | Resultados/Documento |

## Literature Review

### Trabajos Previos

Dentro de la literatura de la industria área, se aborda la problemática de *Dynamic Pricing* con diferentes soluciones, aproximaciones y enfoques al problema. Desde el enfoque determinístico, las investigaciones más relevantes se aproximan al problema desde dos perspectivas. Por un lado, existe la posibilidad de estimar las elasticidades de los precios con los que se puede determinar la respuesta de la demanda a las fluctuaciones de los precios. Por otro lado, está la opción de pronosticar la demanda latente y estimar la sensibilidad al precio del cliente.

(Mumbower, Garrow, & Higgins, 2014) estiman las elasticidades de los precios como un primer paso para integrar la demanda, el pricing y la optimización de ingresos. Esta investigación se basa en información de 21 fechas de vuelo recolectada por *Automated Web Client Robots* (o *webbots*) donde se obtiene información por vuelo acerca de las tarifas y el *seat map* en un horizonte de tiempo de 28 días antes de la salida del vuelo. Adicionalmente, utilizan un modelo de regresión lineal, el cual utilizando una estimación de mínimos cuadrados se detecta que existe endogeneidad en el modelo y por lo tanto los estimadores serán sesgados. Dado lo anterior, se implementa una metodología de dos etapas de mínimos cuadrados (2SLS) con variables instrumentales (Greene, 2002).

Etapa 1

Etapa 2

Las variables exógenas e instrumentales que se tomaron en el modelo fueron las siguientes:

|  |  |
| --- | --- |
| **Variable** | **Descripcion** |
|  | Número total de reservas diarias para un vuelo (variable dependiente) |
|  | Precio promedio del vuelo |
|  | Indica la fecha en que la competencia estaba ofreciendo una tarifa promocional |
|  | Indica reservas realizadas para viajar en o alrededor del *Labor Day* |
|  | Indica la hora de salida del vuelo |
|  | Indica una reserva realizada 1,…,28 días antes de la salida del vuelo |
|  | Indica si un vuelo sale en Lunes,…,Domingo |
|  | Indica si una reserva se realizó en Lunes,…,Domingo |
|  | Variable “Dummy” para cada mercado |
|  | Precio promedio de otros mercados |
|  | Promedio de vuelos de la competencia en un mercado |

En otra investigación realizada por (Granados, Gupta, & Kauffman , 2012) se analiza el impacto del Internet en la demanda, comparando las funciones de demanda de los diferentes canales de venta (Online y Offline). A pesar de que el enfoque del estudio es diferente del propósito de esta investigación, es importante tener en cuenta la metodología que se utilizó para obtener las elasticidades de los diferentes canales. En esta se analizan las elasticidades de precio en los canales online y offline al nivel de industria usando la base de datos de los tiquetes vendidos por agencias de viaje. Se considera un modelo de demanda lineal y log lineal, de tal manera que donde la demanda se estima en términos de cantidad vendida y el precio es el precio promedio en dólares de los tiquetes vendidos para un par de ciudades, canal, segmento y temporada dados. Se concluyó que el modelo log lineal tiene un más alto que el modelo lineal y concluyeron que era el más adecuado a utilizar. A continuación se presenta el modelo log lineal implementado:

Después de definir el modelo se realizó un test de Hausman y se comprobó que existe endogeneidad en las variables por lo que se determinó que un modelo de dos etapas de mínimos cuadrados (2SLS) es el más conveniente para corregir el problema, con las siguientes variables instrumentales:

|  |  |
| --- | --- |
| Variable Instrumental | Descripción |
|  | Distancia en millas entre un par de ciudades |
|  | Grado de concentración del mercado entre un par de ciudades. Se utiliza el índice de Herfindahl |
|  | Operaciones en Hub son asociadas con mayores costos |

(Christ, 2011) implementa un modelo de Dynamic Pricing basado en información de las solicitudes de tarifas que llegan al sistema *Computer Reservation System* (CRS). El enfoque realiza una separación entre la demanda latente como una forma pura de llegada de clientes a partir de la demanda teniendo en cuenta la política de precios prevalente en conjunción con los precios de la competencia y el comportamiento especifico del cliente.

La demanda dado un precio es calculada como la fracción de clientes que actualmente compra el tiquete y la demanda latente , representado de la siguiente manera . La fracion de clientes que tambien es conocida como el *book-to-look ratio* es una función de factores internos y externos relacionados con el precio propio y de la competencia, la demanda latente se asume como una función de factores internos y externos no relacionados con el precio, dando como resultado que

### Two-Stage Leats Squares

En una regresión lineal estándar se asume que los errores de la variable dependiente no están correlacionados con alguna de las variables independientes, en el caso que esto llegue a pasar se dice que existe endogeneidad en el modelo. En el caso del problema tratado en este documento, en previas investigaciones se ha llegado a demostrar que la demanda aérea está influenciada por el precio del tiquete, pero el precio del tiquete está correlacionado con los errores de un modelo de regresión lineal estimado con OLS. El método de dos etapas de mínimos cuadrados permite corregir este problema basado en variables instrumentales (Greene, 2002).

Inicialmente se considera un sistema lineal estándar de ecuaciones, en que la ecuación es de la forma . Donde es un vector de variables independientes, es un vector de variables exógenas, es un vector de coeficientes y es un vector de términos de perturbación de la ecuación . Al definir que los regresores de una o más ecuaciones están correlacionados con las perturbaciones se utilizan variables instrumentales para cada ecuación que pueden ser diferentes o iguales para cada una de las ecuaciones. Estas deben estar correlacionadas con los términos de perturbación de la ecuación . Se estima el modelo con las siguientes dos etapas:

**Etapas**

1. La regresión se obtiene de
2. La regresión ajustada es sustituida por los regresores originales en la ecuación para obtener un 2SLS insesgado donde se estima .

### Variables Instrumentales

Las variables instrumentales son aquellas variables que están correlacionadas con alguna de las variables exógenas pero no pertenecen a la ecuación de la regresión. Para este caso en particular es necesario que las variables instrumentales estén correlacionadas con el precio de los tiquetes pero no con la demanda.

De acuerdo a (Mumbower, Garrow, & Higgins, 2014) existen cuatro tipos de variables instrumentales:

#### Cost-shifting variables

Estas variables generan el desplazamiento de los costos del producto y no estas relacionadas con la demanda. Por ejemplo, (Hsiao, 2008) utiliza las distancias de las rutas multiplicado por el costo de unidad del jet fuel como instrumentos, teniendo que a una mayor distancia mayores costos del jet fuel, por ende, un mayor costo del tiquete. Estas variables no son muy apropiadas para este análisis puesto que no capturan las fluctuaciones del día a día de las tarifas aéreas.

#### Hausman-type Price instruments

Estos instrumentos se basan en la teoría económica de que el precio de una firma en un mercado está en función de los costos marginales promedio de un producto más una cantidad que la firma está dispuesta a cobrar dado el *willigness-to-pay* por un producto en ese mercado específico. Idealmente el precio en un mercado no estará correlacionado con el precio en otros mercados. El precio de un mercado está definido como el precio promedio de todos los otros mercados de trayectos similares.

#### Measures of Competition and Market power as instruments

Esta variable utiliza a la competencia como instrumento. Por ejemplo, en el modelo que desarrollan (Mumbower, Garrow, & Higgins, 2014) utilizan el número de vuelos de competidores en el mercado o número de competidores en un mercado.

#### Non-price product characteristics of other products as instruments

Este instrumento utiliza las características del producto de una firma como variable. La capacidad promedio de otros vuelos operados por la aerolínea en el mismo mercado, pero generalmente no son valores que fluctúen diariamente, por lo que tampoco es apropiado.

## Data Understanding

Actualmente, Avianca Holdings cuenta con diversas fuentes de información que se basan en los diferentes datos que se obtienen diariamente de su operación. Un claro ejemplo de esto, son el sinnúmero de transacciones que se realizan regularmente, entre las que se encuentran las reservas de tiquetes, la expedición de tiquetes, entre otras. Otro ejemplo es la operación de Avianca Holdings con los diferentes vuelos que se operan día a día o la información de vuelos de diferentes aerolíneas.

Dadas las diferentes fuentes de información se definió que la información más adecuada para el modelo tiene que reflejar la opción que escogió (compró) el cliente de las diferentes opciones que se ofrecen en cuanto al Origen Destino, el producto y la anticipación de compra sin tener en cuenta si al final el cliente utilizo el tiquete, esto con el claro propósito de reflejar la intención de compra de los clientes. Por otro lado, es relevante incluir información complementaria, como la cantidad de sillas que está ofreciendo Avianca Holdings en un momento dado para una ruta dada, así como, las frecuencias y la cantidad de sillas que ofrece la competencia.

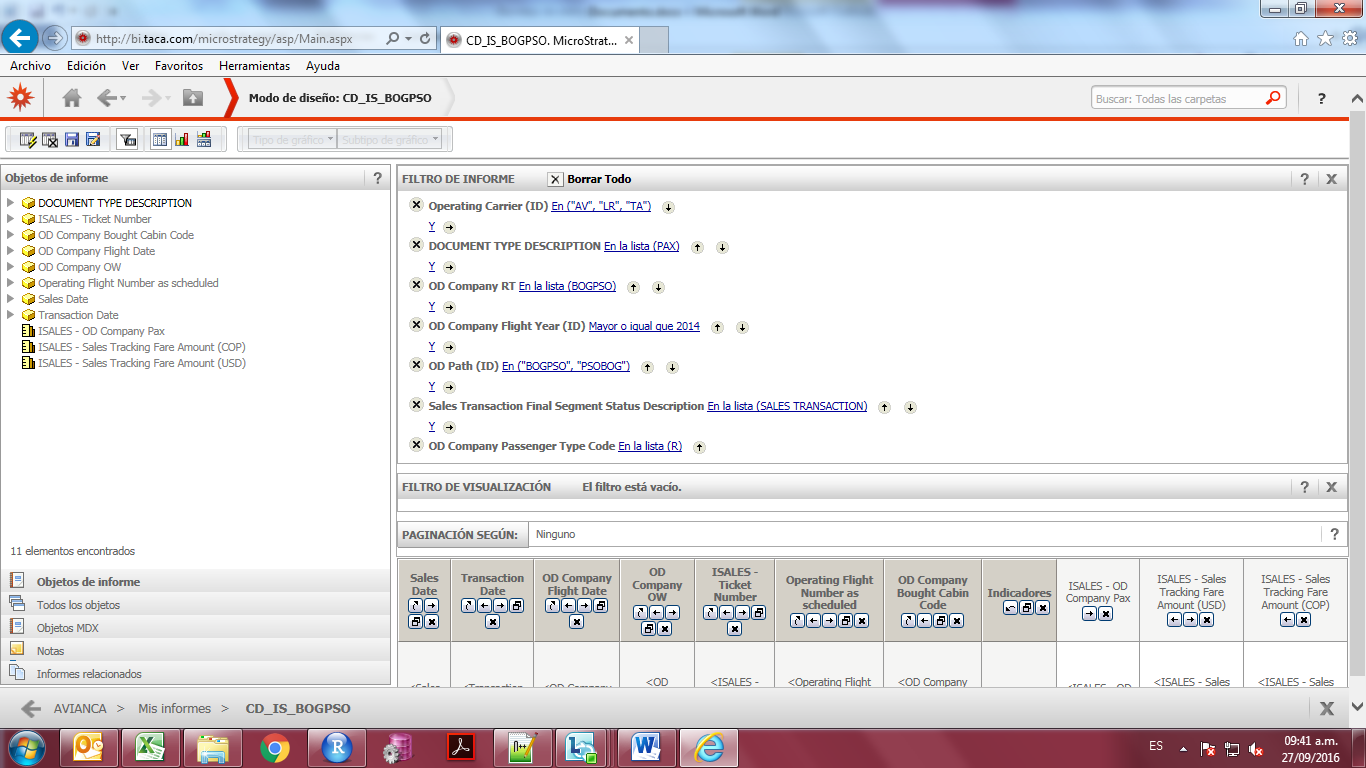
### Describe Data

#### Tiquetes vendidos

Existen tres fuentes principales de las que se pueden obtener los tiquetes comprados por los clientes en base a las características de los tiquetes y a un precio dado para la red de Avianca Holdings. Estas fuentes de información son:

* Flown
  + Información sobre el ingreso de Avianca Holdings que se genera en concepto del servicio de transporte aéreo de pasajeros. Esta información solo incluye a los cupones utilizados y transportados en cualquiera de los diferentes vuelos que opera Avianca
  + No es la fuente más apropiada ya que solo cuenta con los tiquetes utilizados mas no los tiquetes que realmente fueron comprados por los clientes
* TCN
  + Contiene información de los boletos emitidos a través de los sistemas de reservación automatizados y de sistemas propios de ticketing. Esta información es utilizada para hacer análisis de ventas y también es una fuente importante para estimar los ingresos.
* Venta Integrada
  + Integra la información de fuentes relacionadas al tema de venta entre la que están ARC/BSP, E-Star, TCN y Sales RAPID. Esta información tiene tanto boletos de pasajeros como documentos misceláneos.
  + Cuenta con las siguientes ventajas:
    - Se tiene la posibilidad de emular la foto con la fecha de transacción
    - La métrica de *tracking* muestra un comportamiento confiable al sustituir a TCN
    - El atributo de *transaction type* permite evaluar por aparte la venta de las cancelaciones
    - Muestra una intención Origen – Destino más cercana a Flown

Teniendo en cuenta las características de las fuentes, se tomó la decisión de utilizar como fuente de venta de tiquetes a “Venta Integrada” dadas las ventajas mencionadas anteriormente. Se utilizó la siguiente consulta en *Microstrategy* para obtener los datos de los tiquetes:



El set de datos completo obtenido para ejecutar el modelo tiene el siguiente formato:

|  |
| --- |
| **VENTA INTEGRADA** |
| Sales.Date |
| Transaction.Date |
| OD.Company.Flight.Date |
| OD.Company.OW |
| ISALES...Ticket.Number |
| Operating.Flight.Number.as.scheduled |
| OD.Company.Bought.Cabin.Code |
| ISALES...OD.Company.Pax |
| ISALES...Sales.Tracking.Fare.Amount..USD. |
| ISALES...Sales.Tracking.Fare.Amount..COP. |

#### Disponibilidad de Sillas

La fuente más adecuada para obtener la información de la disponibilidad de sillas para cada uno de los vuelos y para cada uno de los días anteriores al vuelo es VQs. Esta fuente de información contiene la capacidad física y ajustada, así como las reservas que se realizan día a día para cada vuelo de la red de Avianca Holdings. El set de datos completo obtenido para ejecutar el modelo tiene el siguiente formato:

|  |
| --- |
| **VQs** |
| OD\_OW |
| FlightDate |
| AdvancePurchase |
| FlightNumber |
| CapacityC |
| CapacityY |
| BookingsC |
| BookingsY |
| AvailabilityC |
| AvailabilityY |

#### Holidays and Special Events

El área de Demanda facilito la información de los festivos y eventos especiales para las rutas más importantes del domestico de Colombia. El set de datos completo obtenido para ejecutar el modelo tiene el siguiente formato:

|  |
| --- |
| **VQs** |
| OD\_OW |
| Date |
| HSE |

#### Departure Time Window

Contiene los *Time of Departure* en los que se agrupan los vuelos de acuerdo a la hora de salida. Esta información se obtiene con los *Departure Time Window* que define PROS para cada *OD Path*. El set de datos completo obtenido para ejecutar el modelo tiene el siguiente formato:

|  |
| --- |
| **TOD** |
| OD\_OW |
| BeginTime |
| EndTime |
| TOD |

#### Competencia

*Industry Schedules* contiene la información de los vuelos de pasajeros y carga itinerados en el pasado y para volar un año a futuro, de todas las aerolíneas disponibles en la base de datos del proveedor OAG. El set de datos completo obtenido para ejecutar el modelo tiene el siguiente formato:

|  |
| --- |
| **INDUSTRY SCHEDULES** |
| FlightDate |
| FlightHour |
| LEG\_OW |
| Seats |

#### Datos

Junto con Demanda se definieron 3 rutas iniciales *Round Trip* para realizar el análisis, estas fueron: BOGPSO, BGABOG y BOGEYP. Se incluyó información desde 2014 para tener información completa de todas las fuentes de información.

### Explore Data and Data Quality

Se seleccionó el OD Local BOGPSO para realizar el análisis exploratorio.

Se inicia verificando la cantidad de tiquetes comprados desde 2014 para cada día de anticipación de compra. Se grafica el siguiente gráfico:

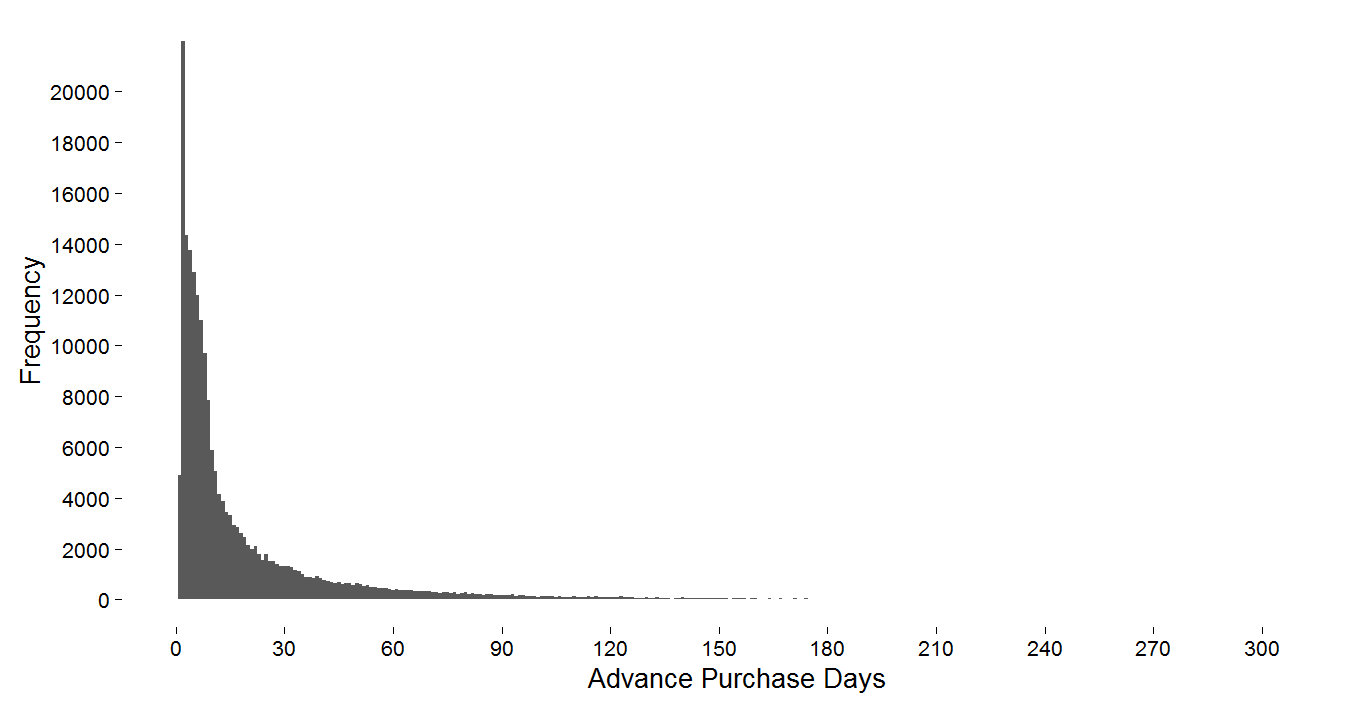


Ilustración 1. Bar graph - Tiquetes vs. Anticipación de Compra

En general, a medida que el tiempo se acerca a la fecha de salida del vuelo se venden más tiquetes. Para el caso específico del OD local BOGPSO, se observa que en los últimos 30 días se concentra la mayoría de la venta y que en los últimos 90 días se venden aproximadamente el 96% de todos los tiquetes vendidos. Por lo tanto, es recomendable hacer énfasis en cierto periodo de venta. Para este caso, se tendrá en cuenta la información de los últimos tres meses para realizar los posteriores análisis y modelos.

Por otro lado, se analiza el *Ticket Fare* para todos los tiquetes vendidos de donde se obtiene el siguiente histograma:

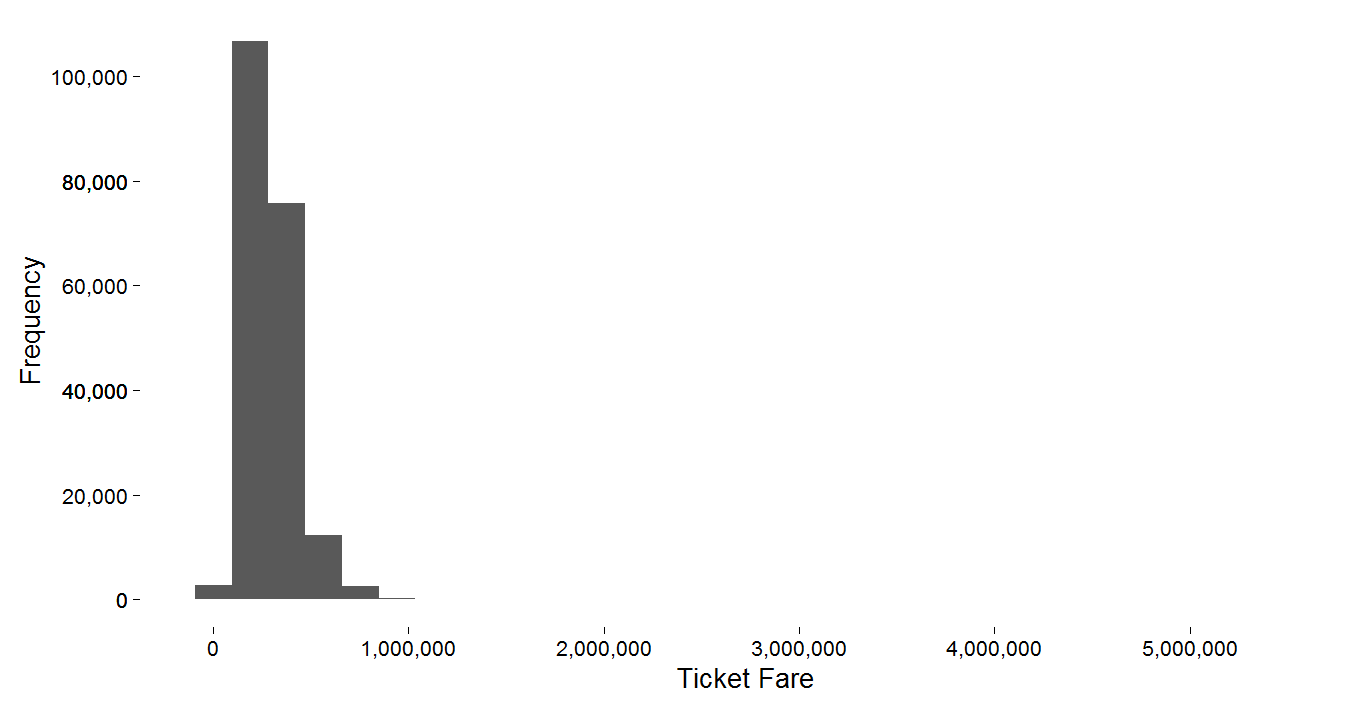


Ilustración 2. Histograma - Valor del Tiquete

Se puede advertir que el *Fare Value* de los tiquetes se concentra entre $0 y $1.000.000 y que existen valores atípicos menores a $10.000 y mayores a $1.000.000. Estos valores generalmente no se ven en el mercado, lo que podrían generar un gran apalancamiento en los siguientes análisis y en el modelo, llevando a un posible sesgo de las estimaciones. Dado lo anterior, es necesario definir un proceso de eliminación de valores atípicos. En este caso, gráficamente se determinó que se eliminaran aquellos valores menos a $10.000 y mayores a $1.000.000. Al filtrar esta información se obtiene que se eliminaron el 0.12% de todos los registros

Dado las anteriores graficas se aplicaron los dos filtros obteniendo una eliminación únicamente del 4.18%. Después de aplicar estos filtros se obtienen las siguientes graficas:

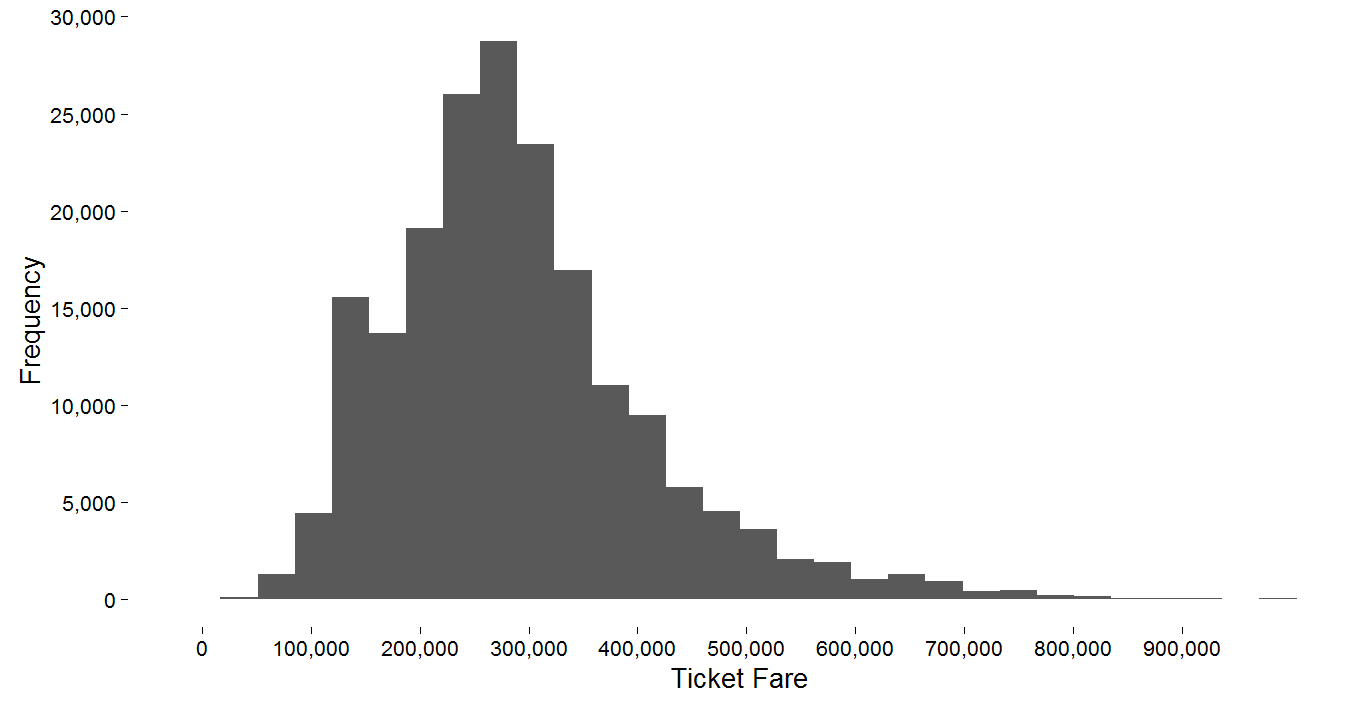
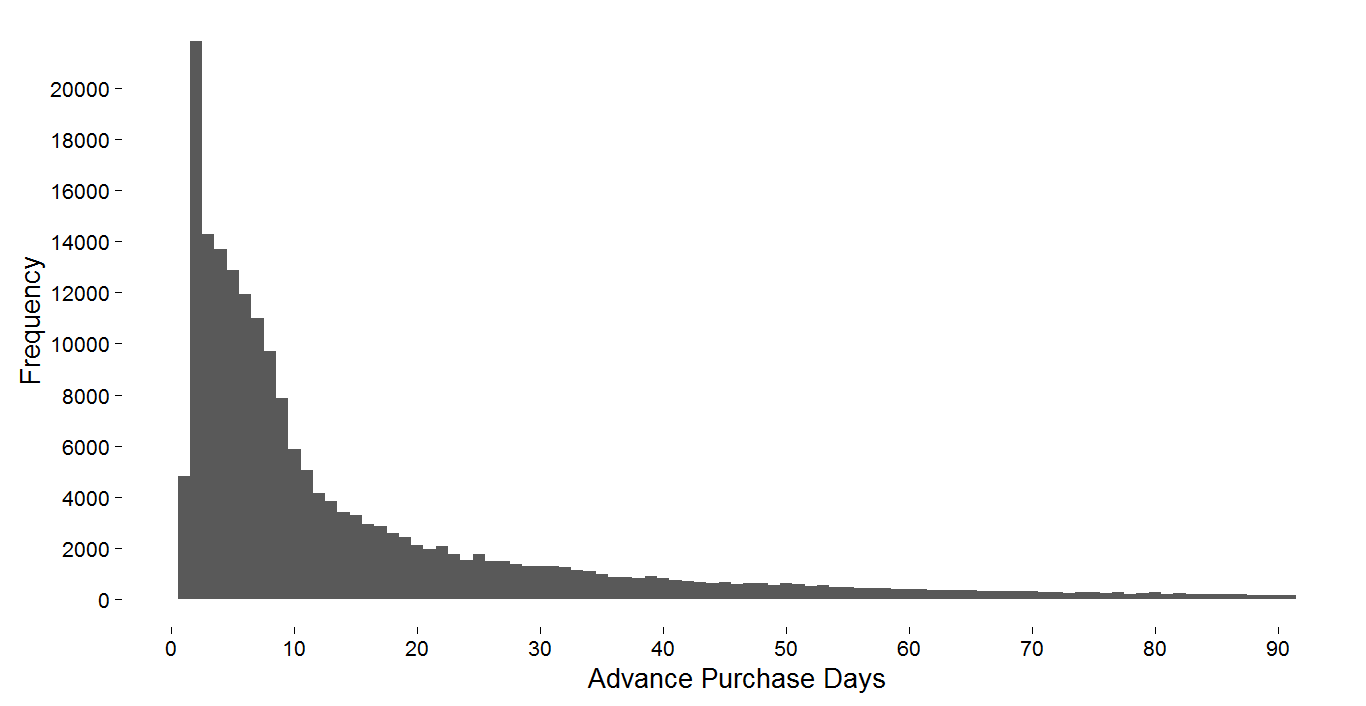


Ilustración 3. Bar graph - Tiquetes vs. Anticipación de Compra (Izquierda), Histograma (Derecha) - Valor del Tiquete

Se agregó la información de días de anticipación de compra a semana y se determina cuantos tiquetes en promedio y desviación de tiquetes son comprados por semana de anticipación de compra.

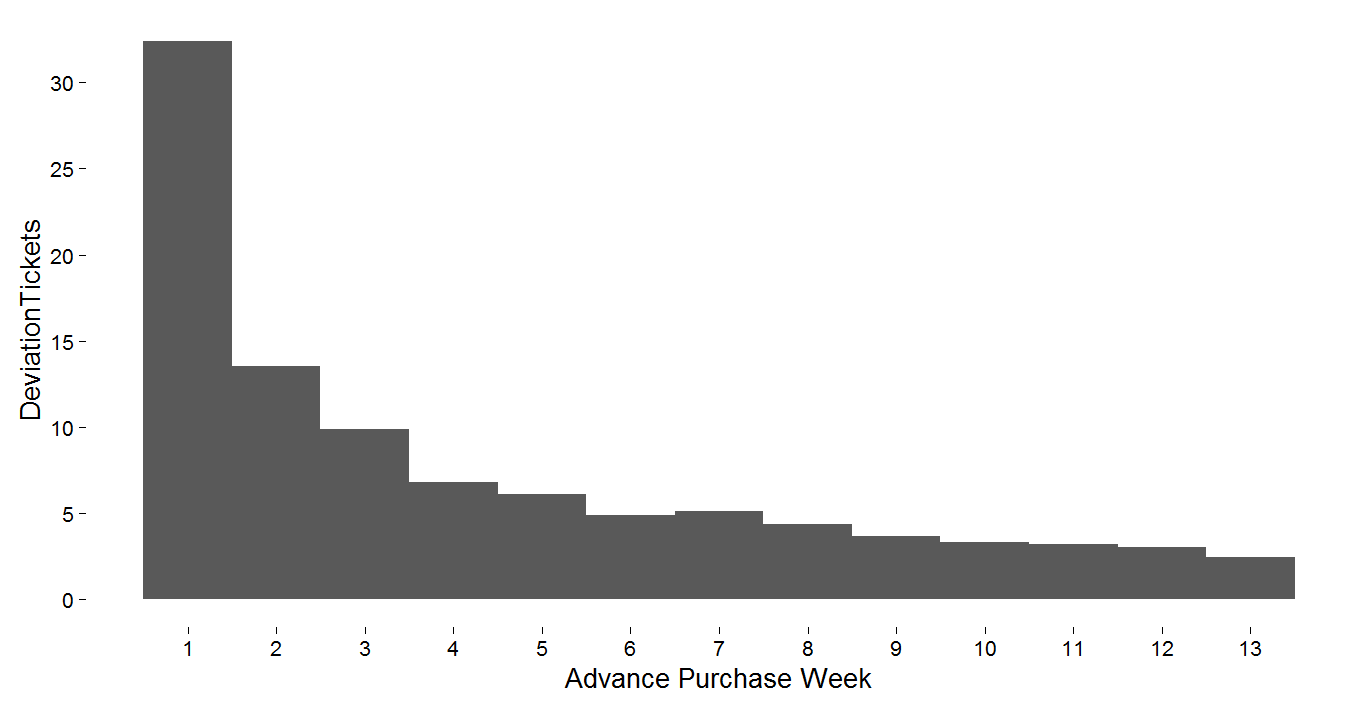
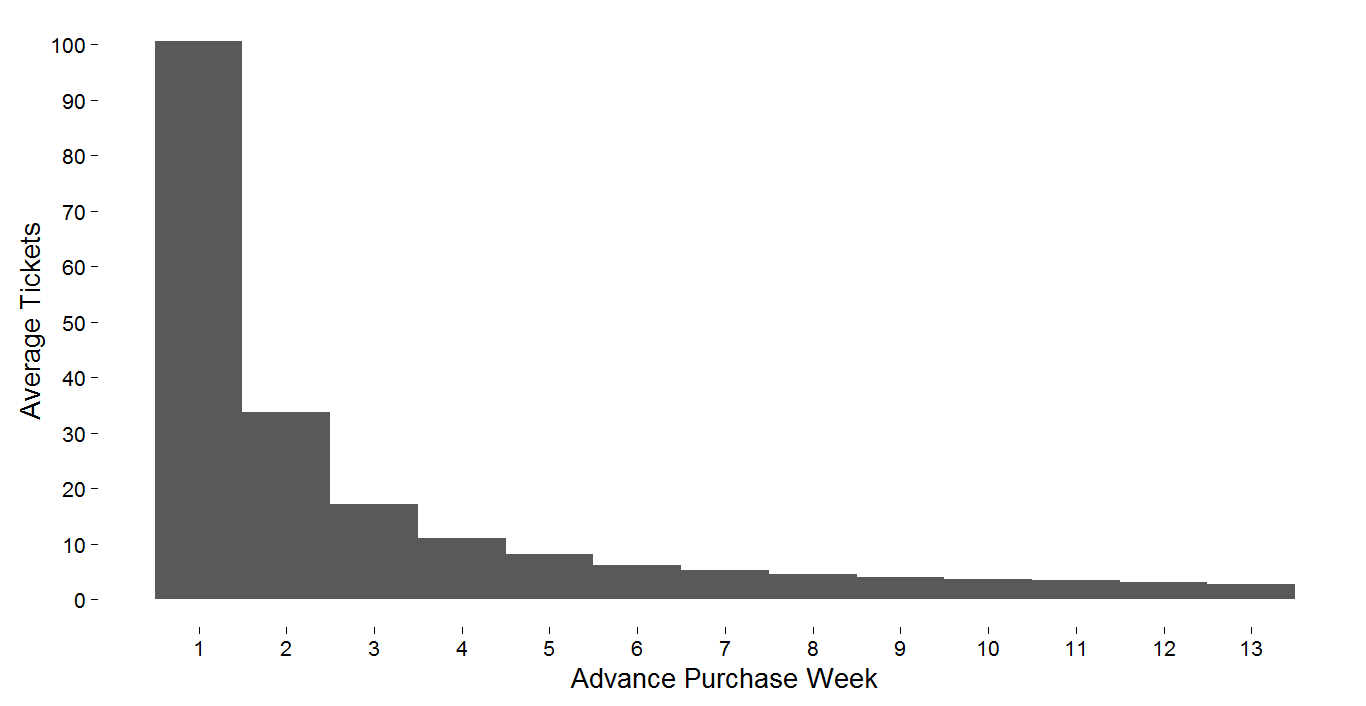


Ilustración 4. Bar graph - Promedio Tiquetes vs. Semana Anticipación de Compra (Izquierda), Bar graph - Desviación Tiquetes vs. Semana Anticipación de Compra (Derecha)

Se visualiza que en la última semana es donde más tiquetes se compran en promedio pero también donde hay mayor variación, por otra parte, las dos graficas tienen un comportamiento similar

Se procede a examinar la variación de tiquetes comprados por día de la semana de venta y día de la semana de vuelo, por medio de los siguientes Boxplots:

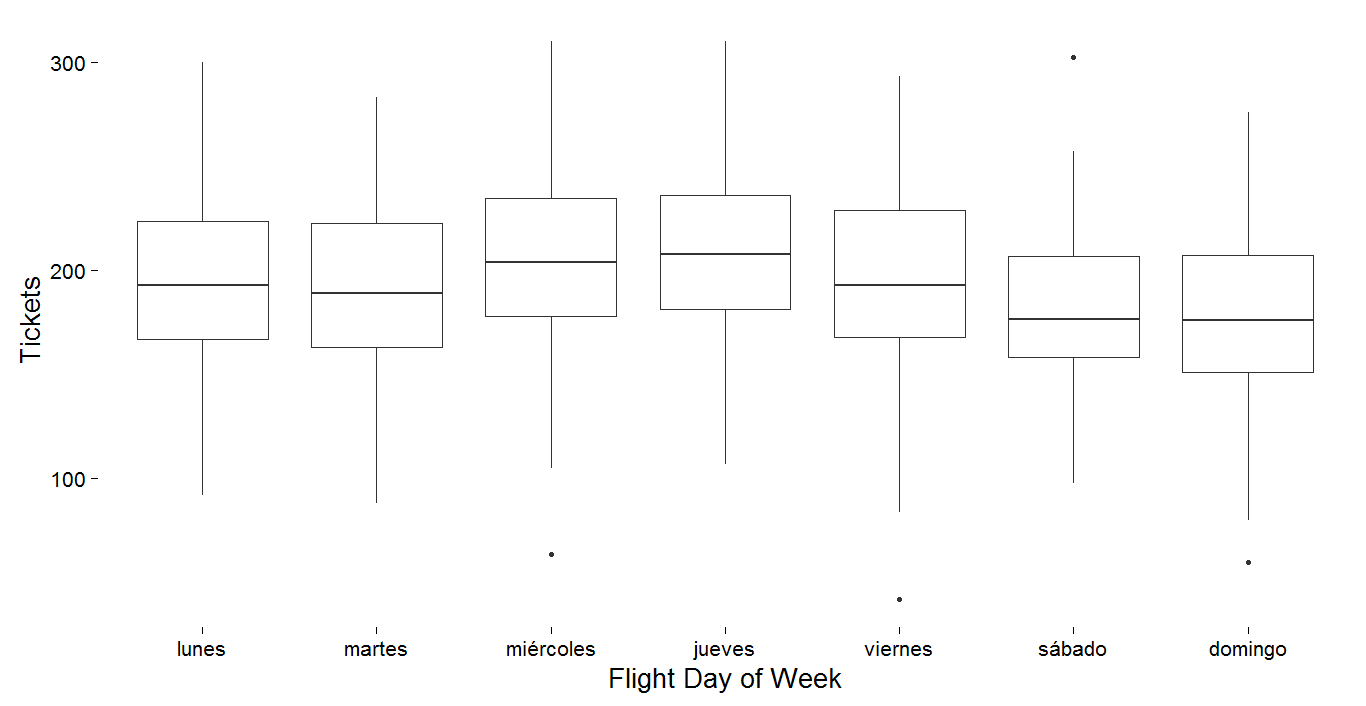
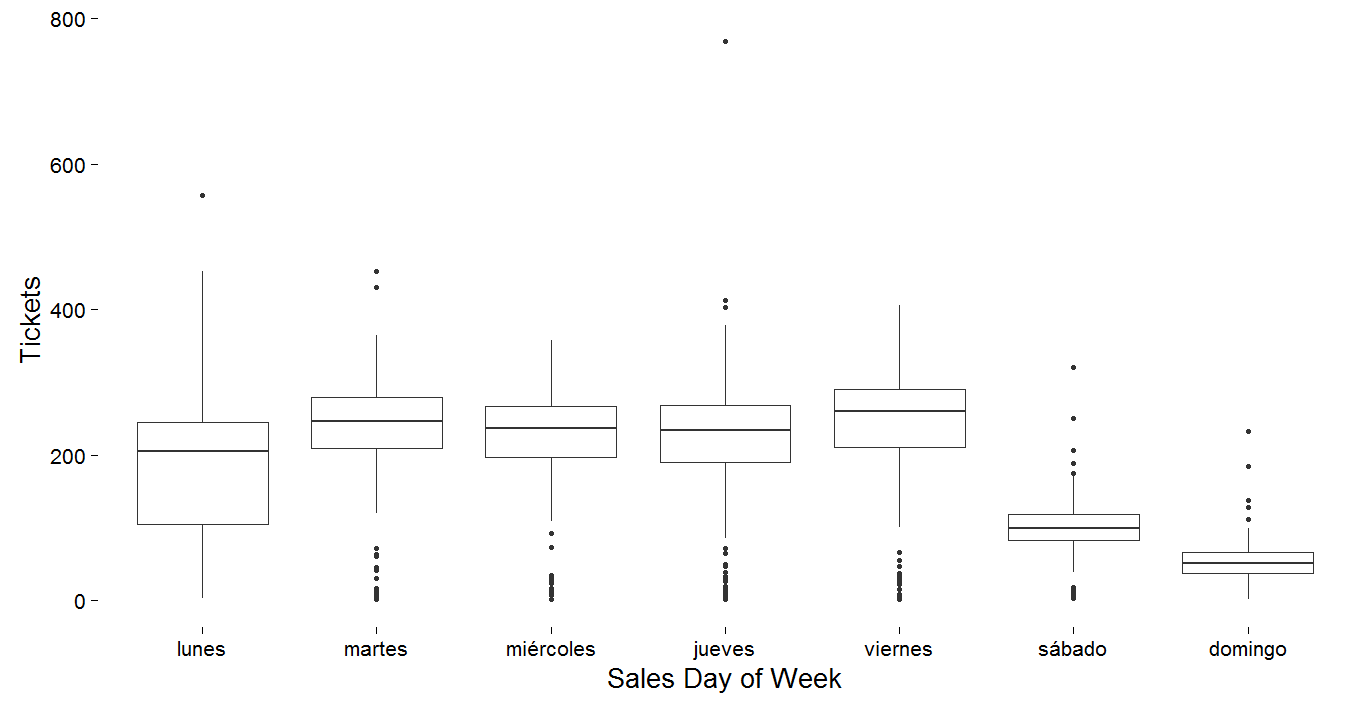


Ilustración 5. Boxplot - Venta Día de la Semana (Izquierda), Vuelo Día de la Semana (Derecha)

Se percibe que el día de la semana de vuelo es relevante para la venta de tiquetes ya que existen diferencias en los días sábado y domingo respecto a los demás días. En estos dos días menos gente compra tiquetes, aunque esto puede estar influenciado a que hay menos cantidad de frecuencias para estos dos días. También se observa que los días de más ventas son los jueves y viernes. En cuanto a los días de semana de venta, es bastante relevante ya que los días del fin de semana es donde menos se vende tiquetes.

Se juntan los días de la semana de venta y los días de la semana de vuelo, con los que se construye el siguiente Heatmap:

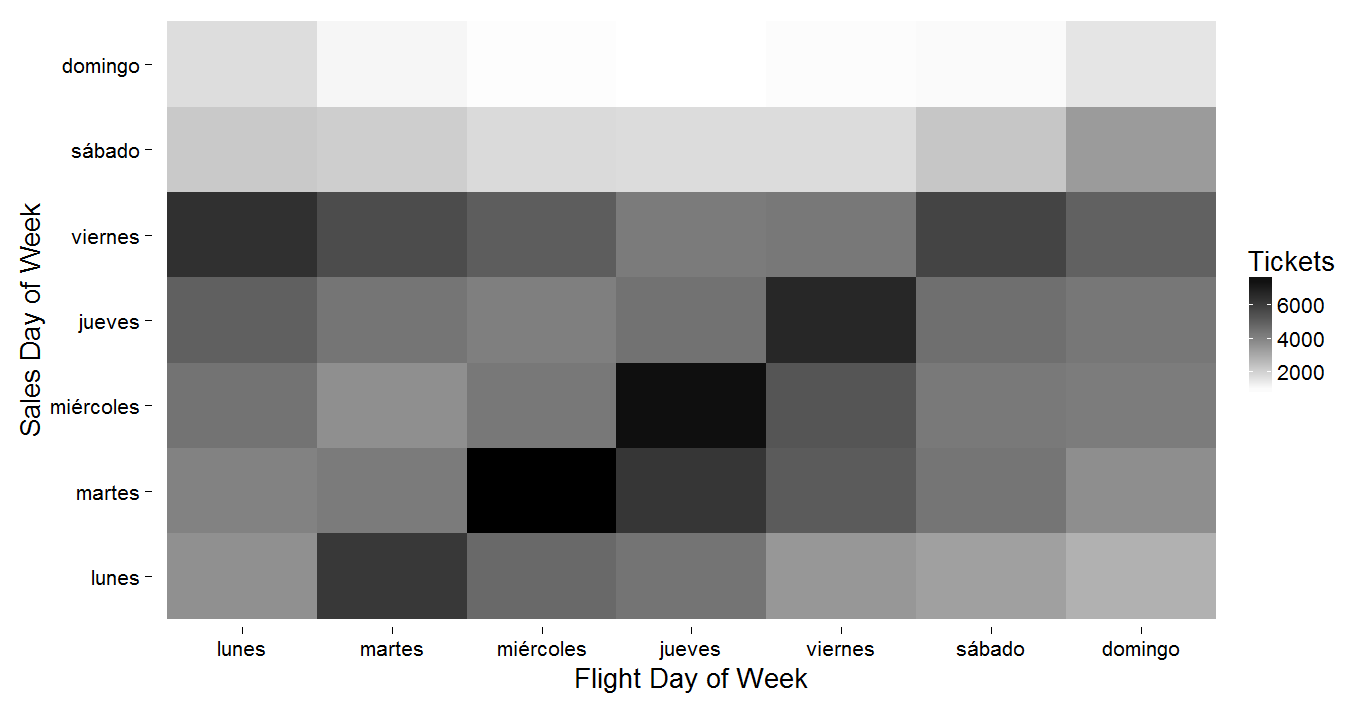


Ilustración 6. Heatmap - Día de la Semana Venta vs. Día de la Semana Vuelo

En este grafico se puede llegar a concluir que muchos de los tiquetes que se compran son para volar el siguiente día, esto basado en la línea vertical que se genera donde existen más tiquetes vendidos. Por otro lado el viernes se compran bastantes tiquetes para viajar los días lunes. Se infiere, que existe una interacción entre estas dos variables.

Posteriormente se evalúa el Departure Time Window de los vuelos que ya operaron y cuál fue el Booking Load Factor que estos tuvieron.

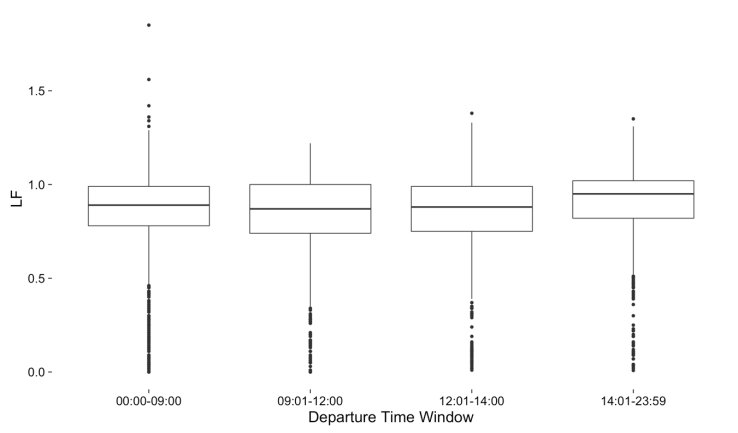


Ilustración 7. Boxplot - TOD

En general, los cuatro TODs que PROS tiene para esta ruta tienen valores similares y no existe una variación significativa entre estos. Para este caso no se tendría que tener una variable con TOD pero para otras rutas donde existen más TODs muy seguramente existirán diferencias en los Booking Load Factor. Esta variable es importante analizarla ya que refleja la demanda que se llegó a tener para cada TOD.

Se evalúan para los diferentes tipos de días que existen en el OD como son los días normales o los días donde existen algún tipo de *Holiday and Special Event.* Se grafican la cantidad de tiquetes vendidos y el el valor de los tiquetes en estas fechas.

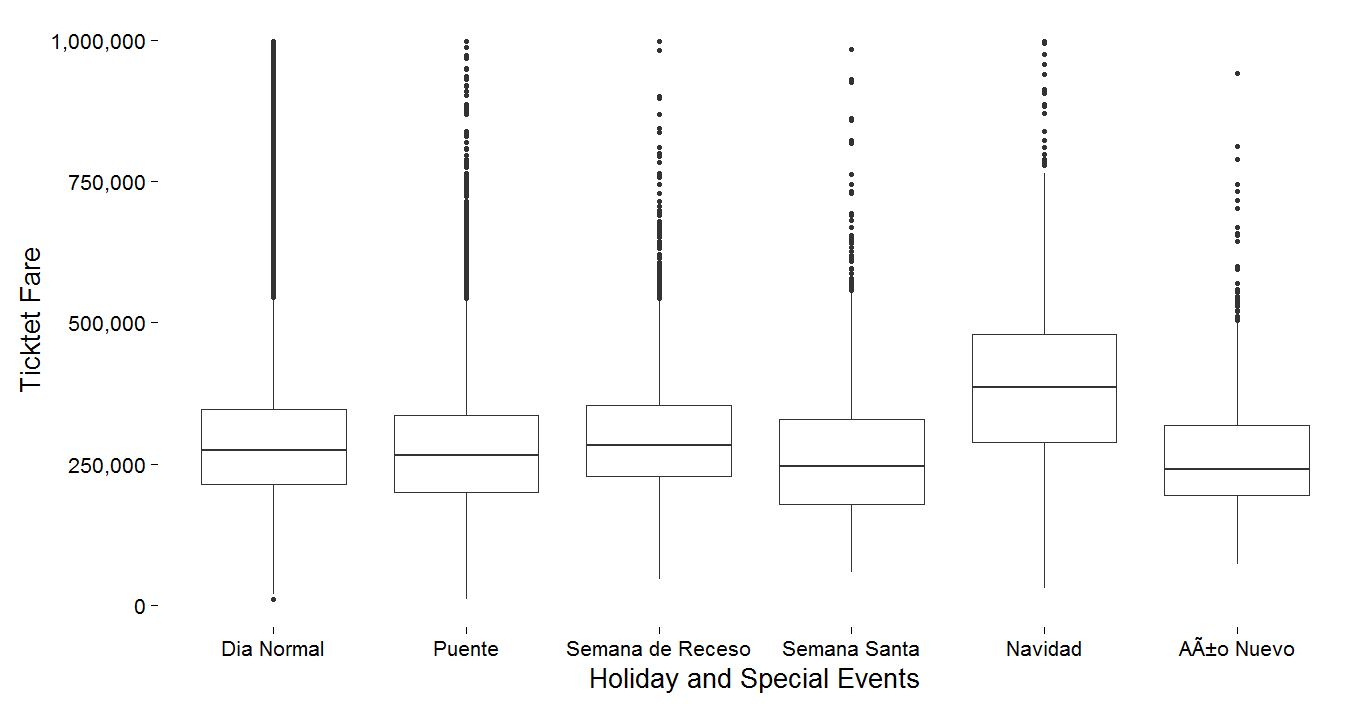
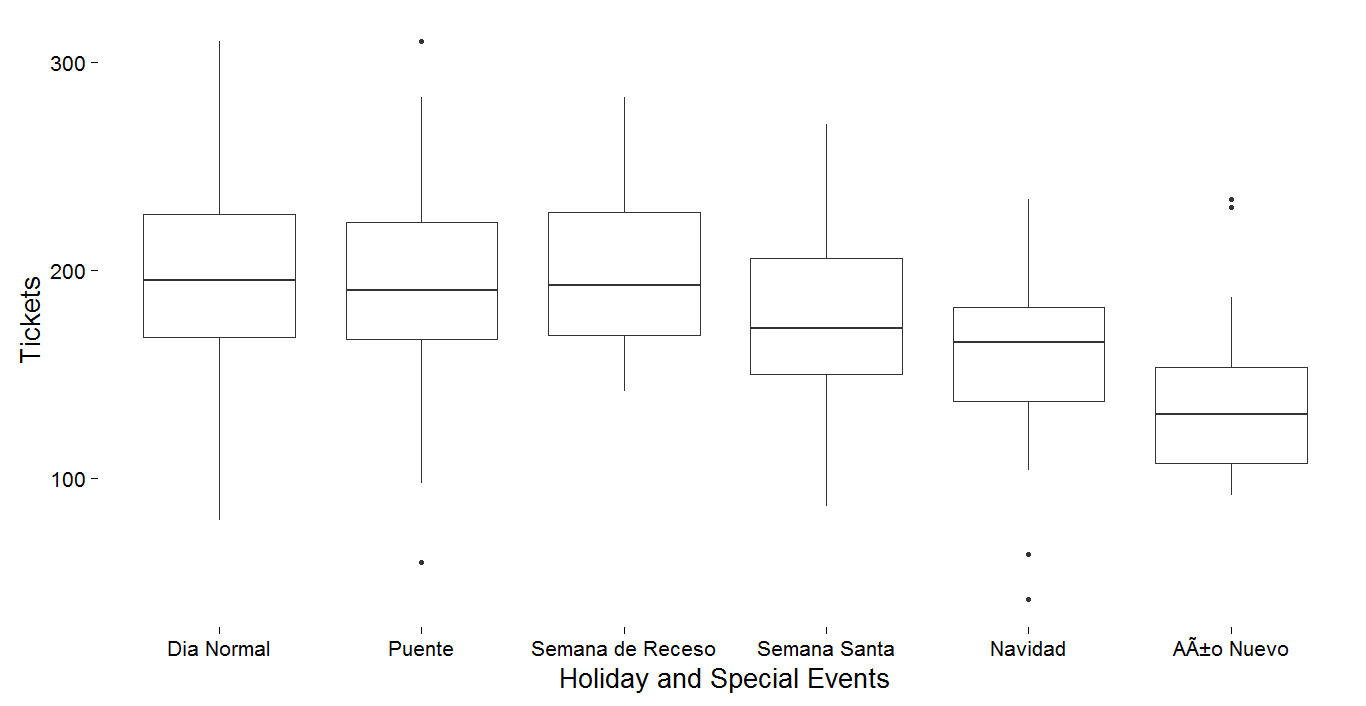


Ilustración 8. Boxplot - Tiquetes Vendidos por Tipo de Día (Izquierda), Tarifa Tiquetes por Tipo de Día (Derecha)

Se observa que para los días normales, puentes y semana de receso se tiene en promedio la misma cantidad de tiquetes y para Año Nuevo la demanda no es tan alta. Por otro lado, el valor de los tiquetes es similar para todas las fechas excepto en navidad donde en promedio se venden tiquetes más caros. Se observa también que existen varios datos atípicos para cada una de las series, valores que hay que llegar a evaluar para no llevar a el sesgo de las estimaciones.

Seguidamente, se grafica el valor del tiquete contra la anticipación de compra y contra la disponibilidades de sillas

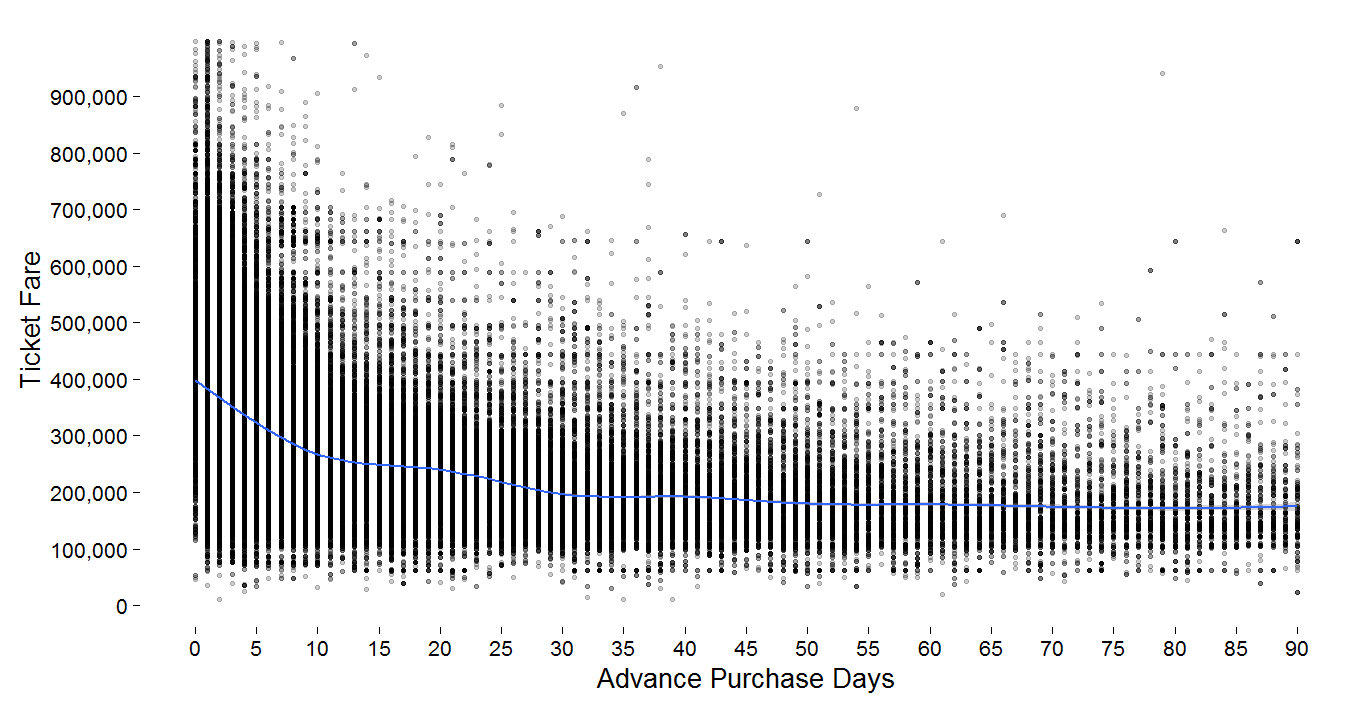


Ilustración 9. Scatterplot - Días Anticipación de Compra vs. Tarifa Tiquete

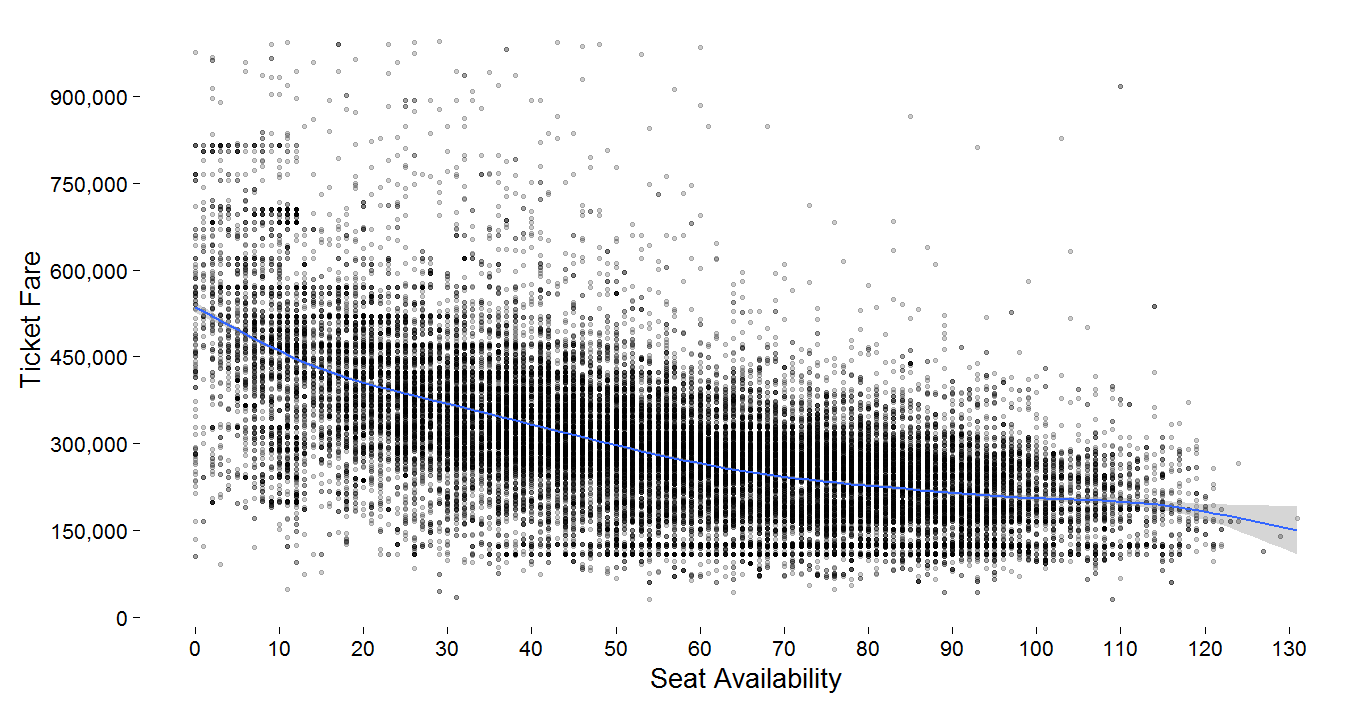
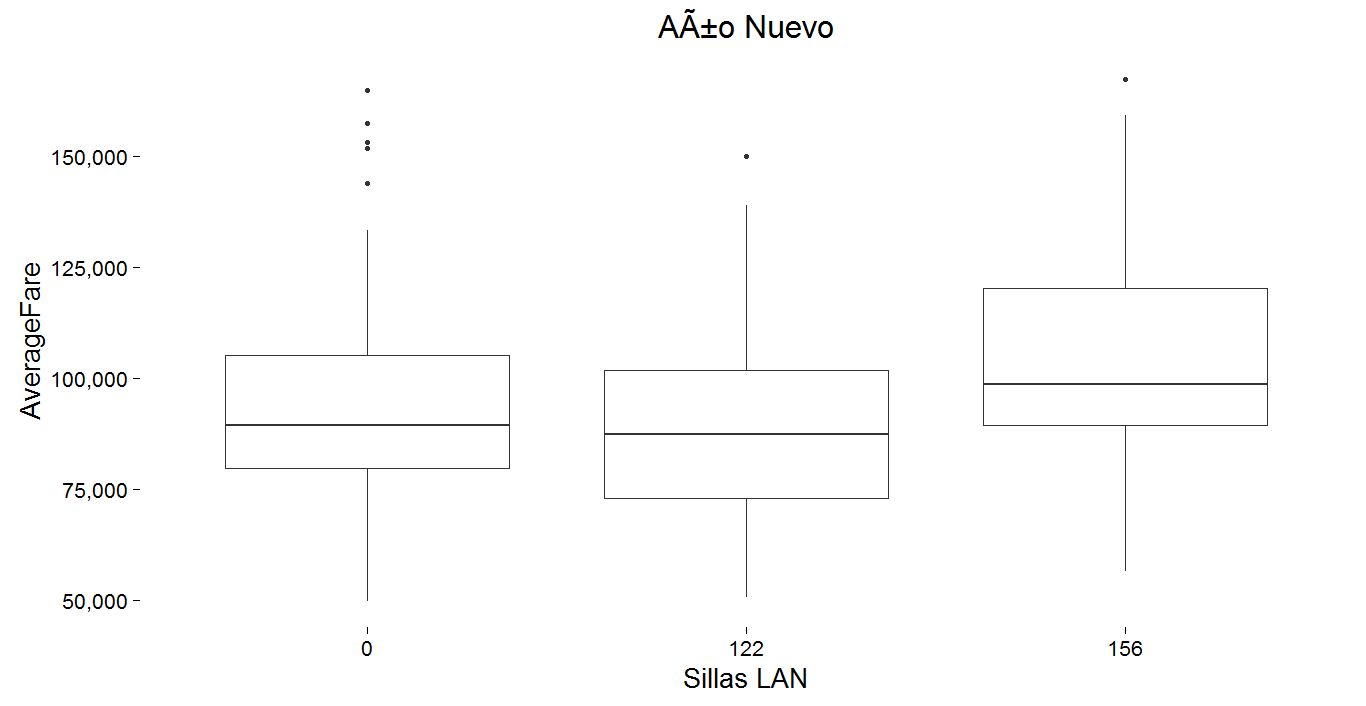
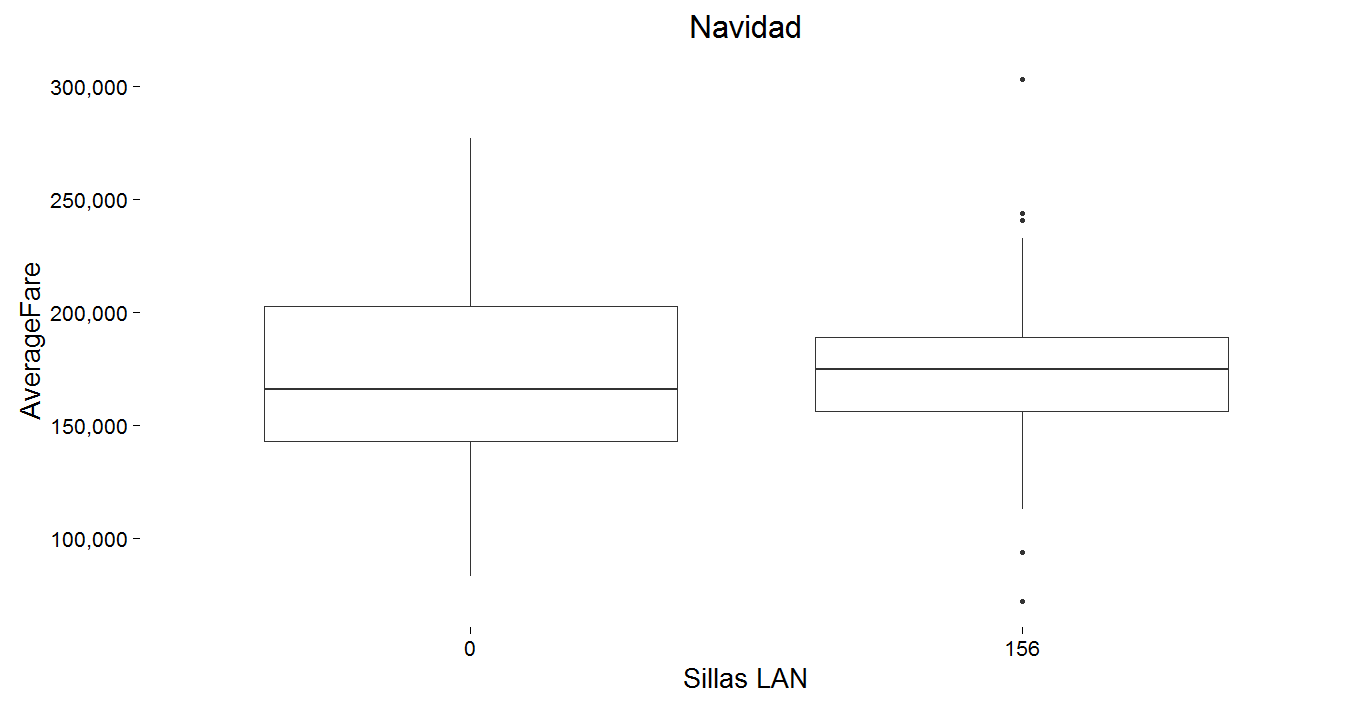
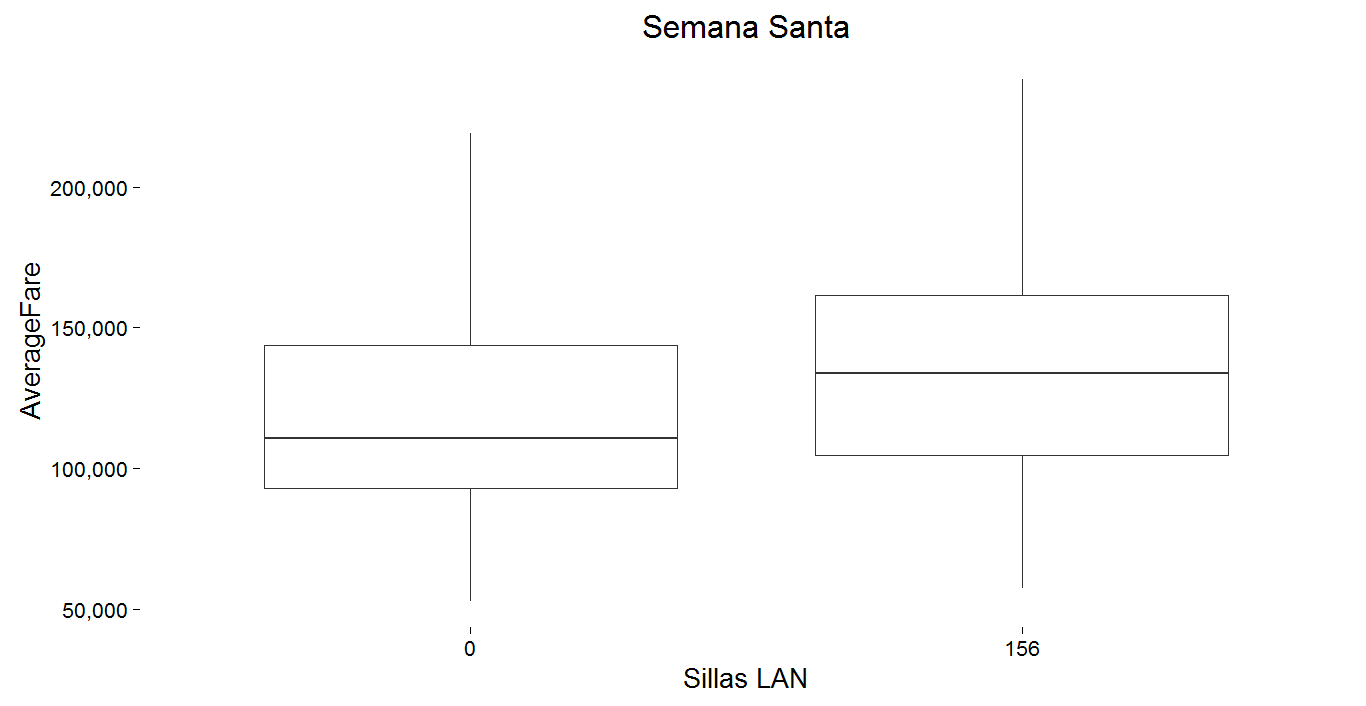
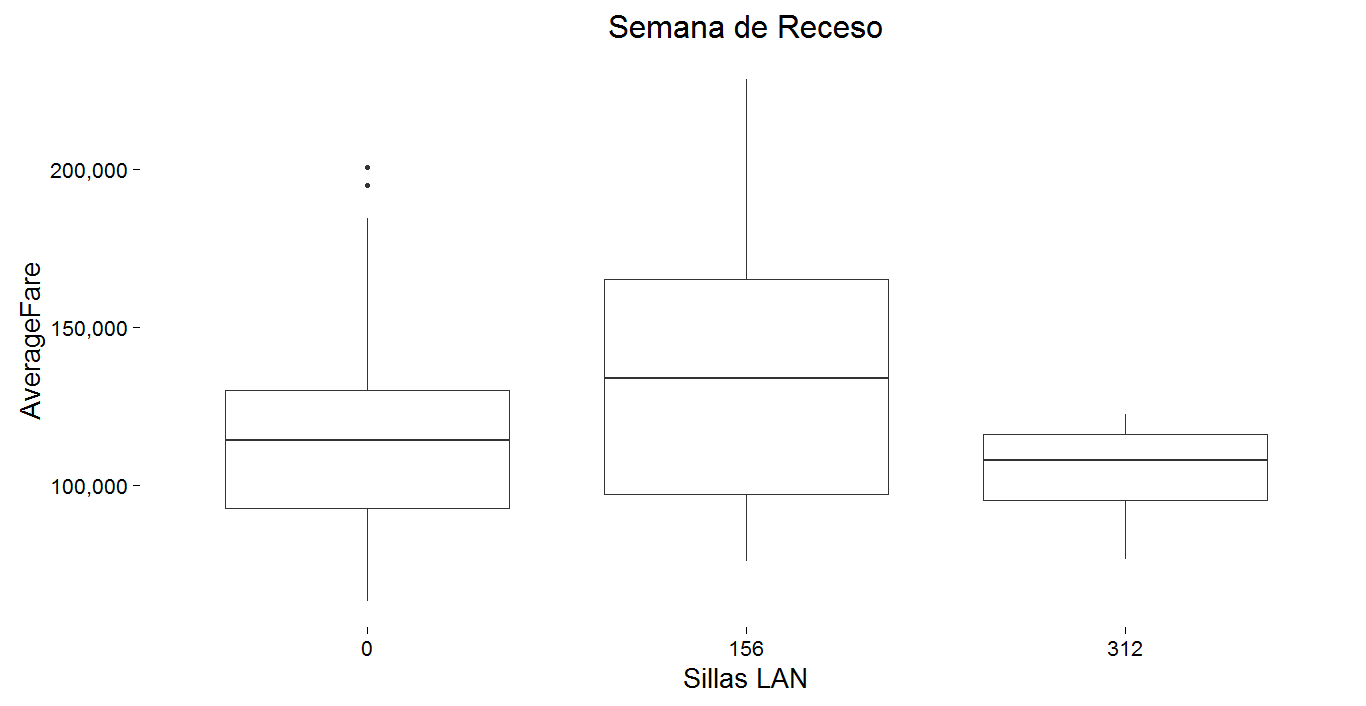
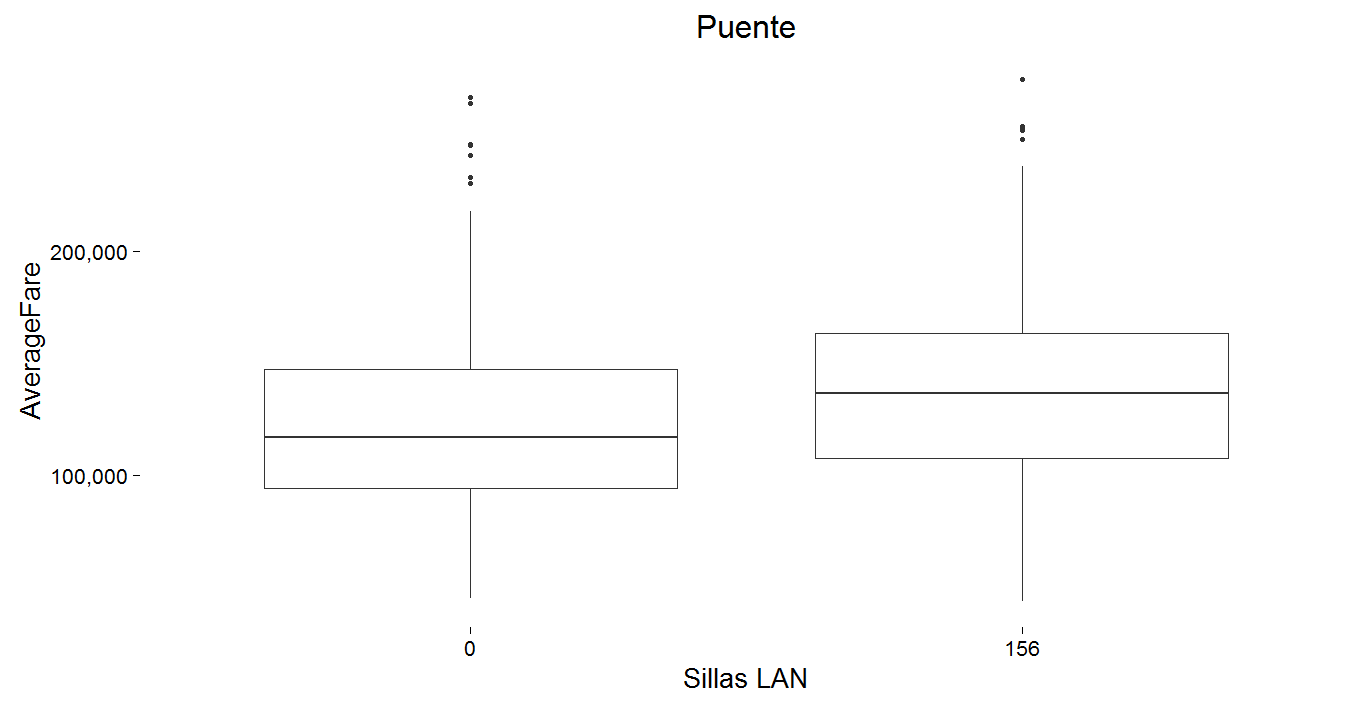
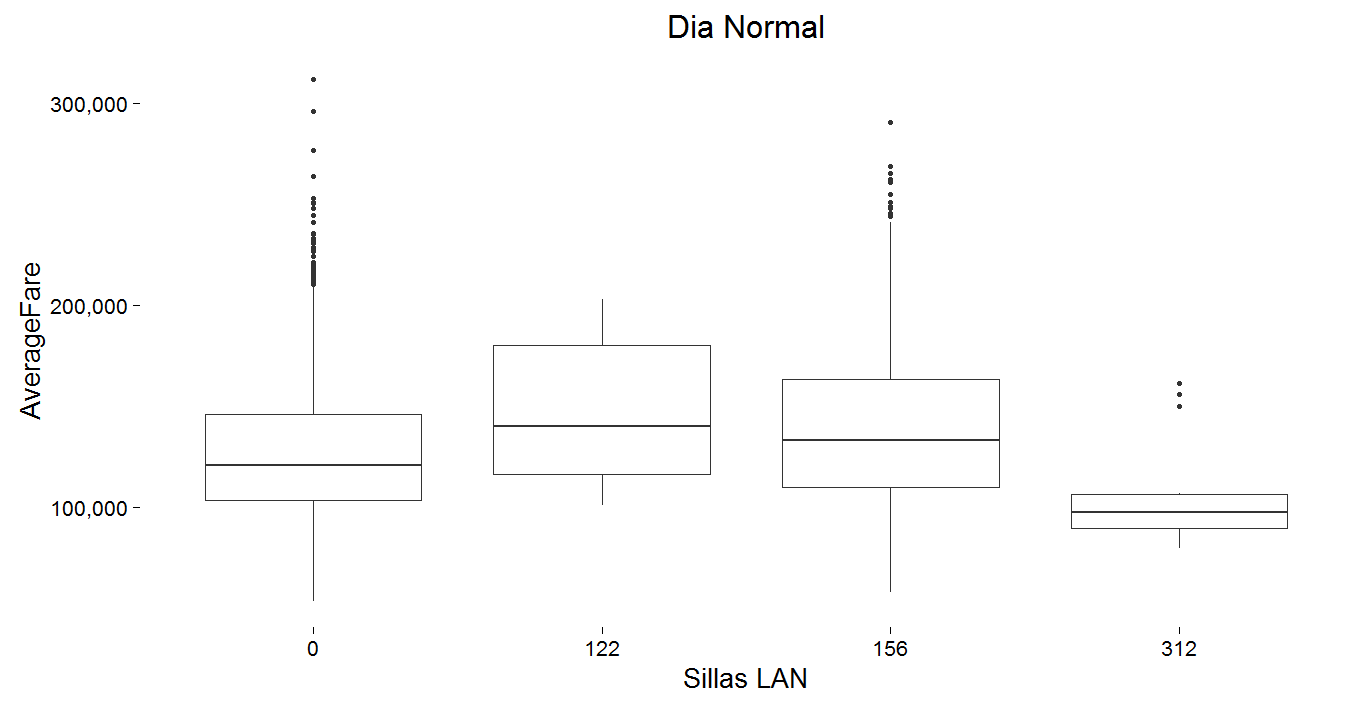


Ilustración 10. Sillas Disponibles vs. Tarifa Tiquete

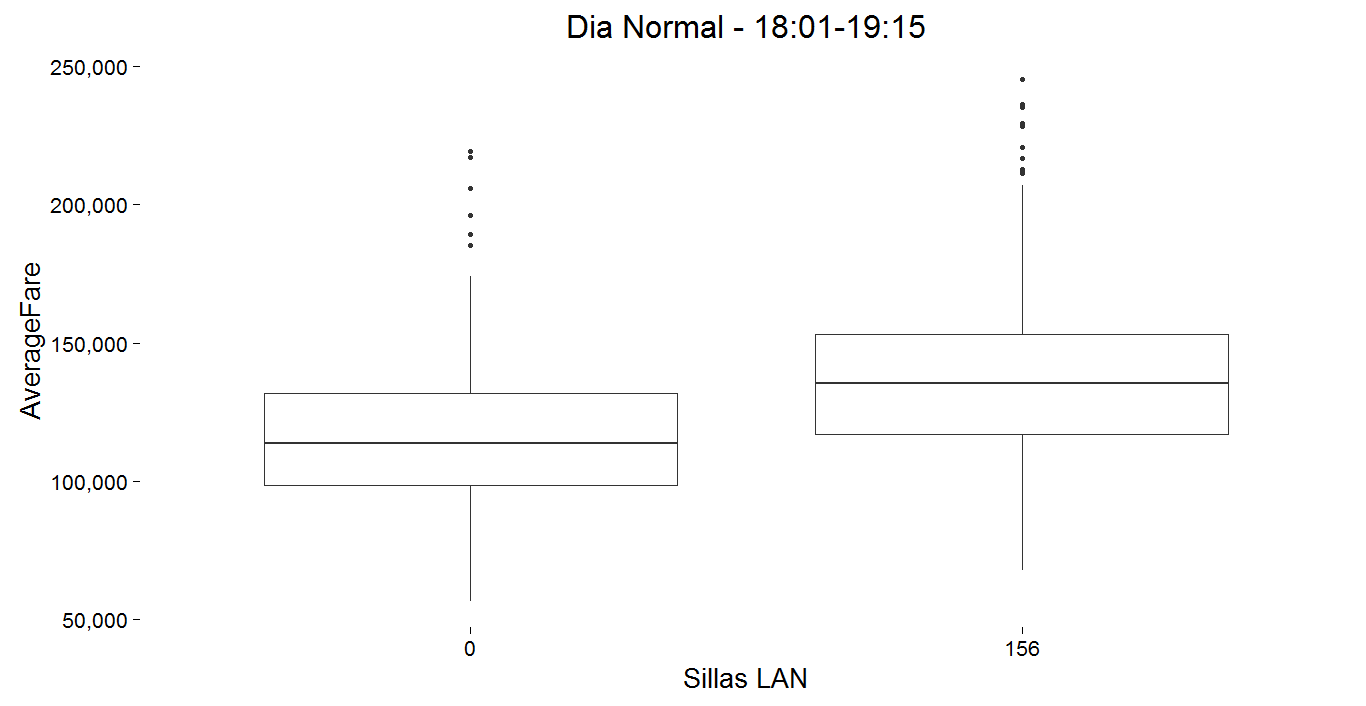
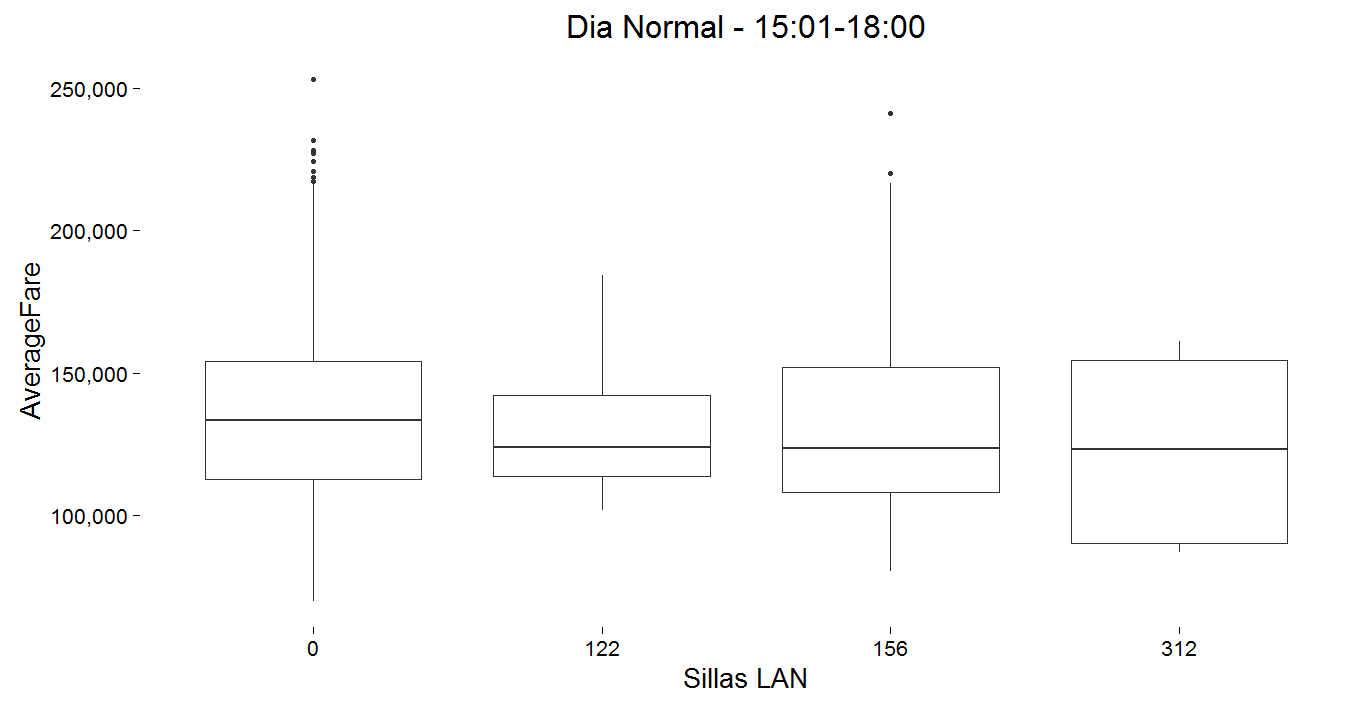
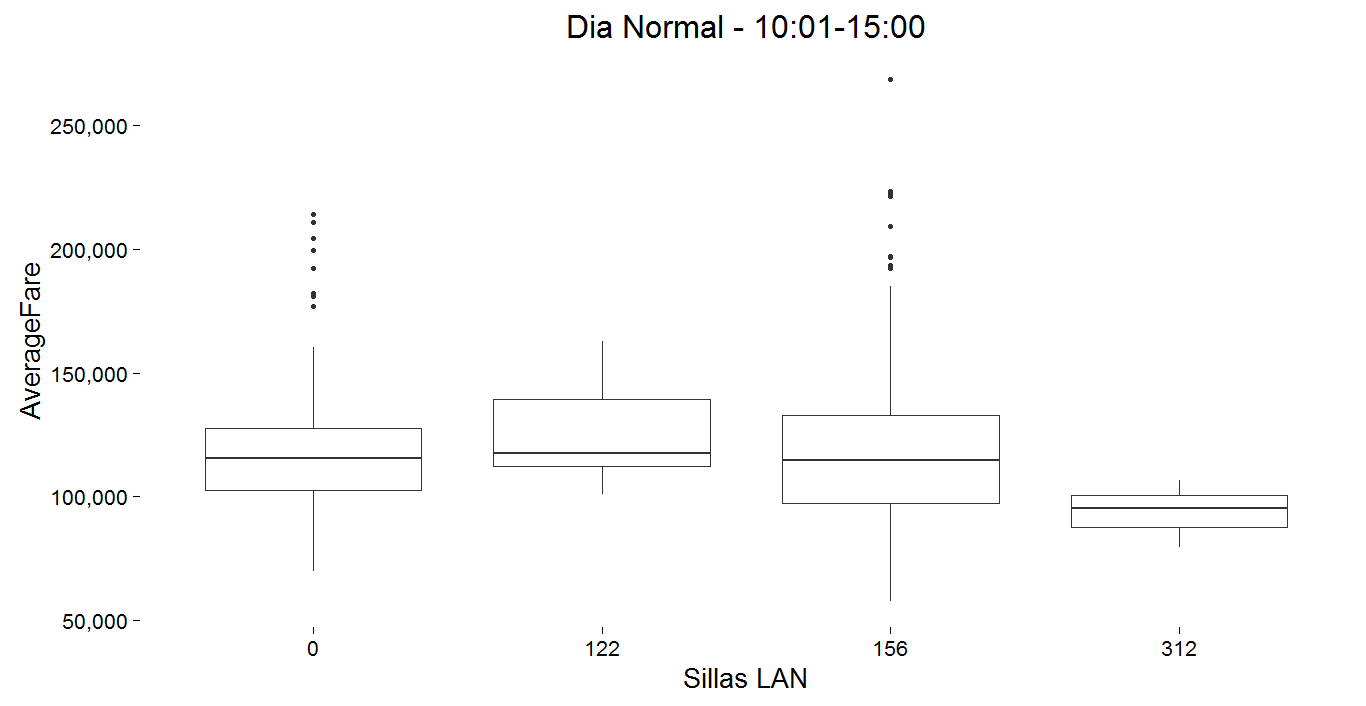
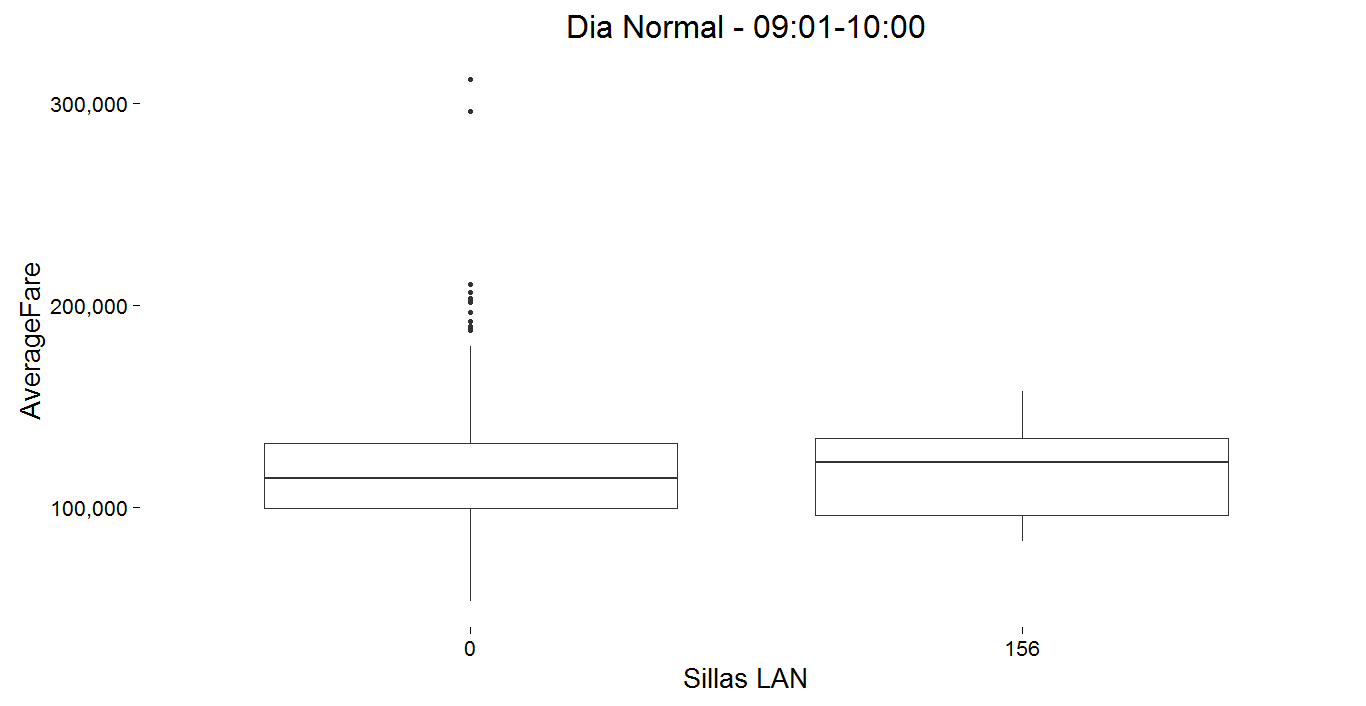
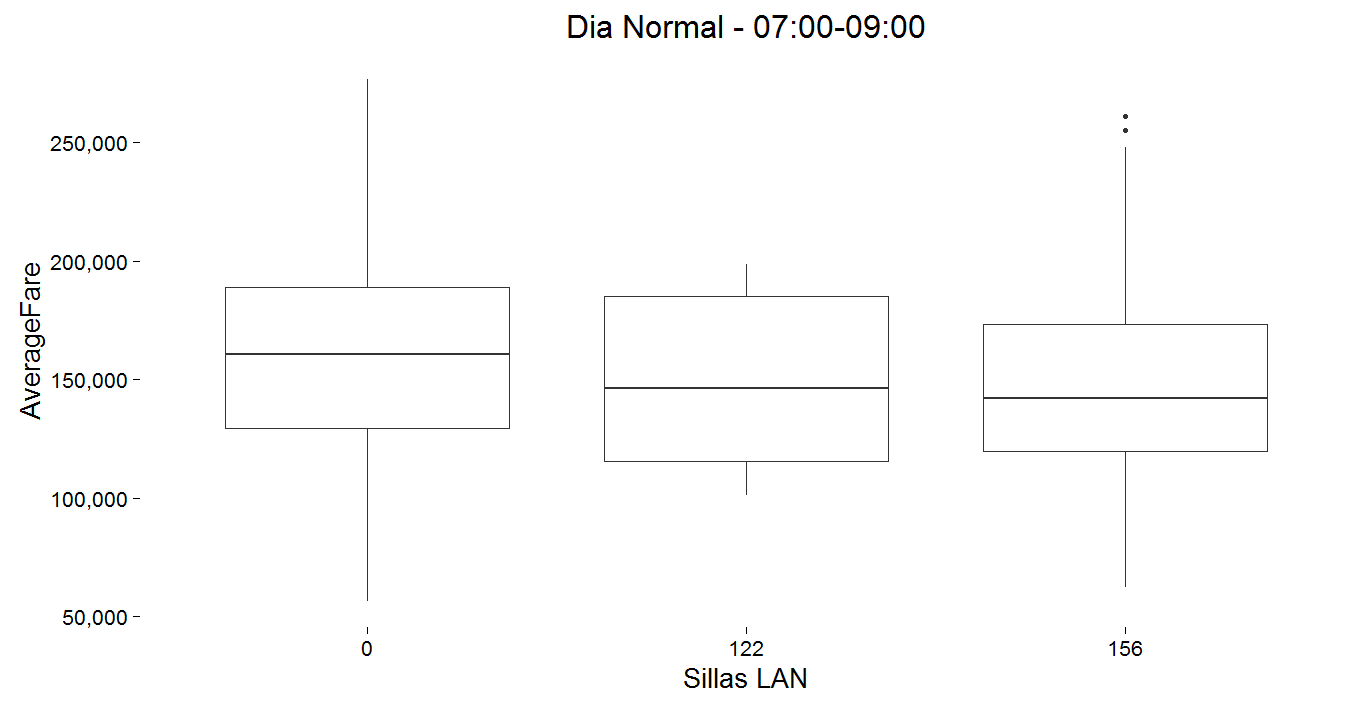
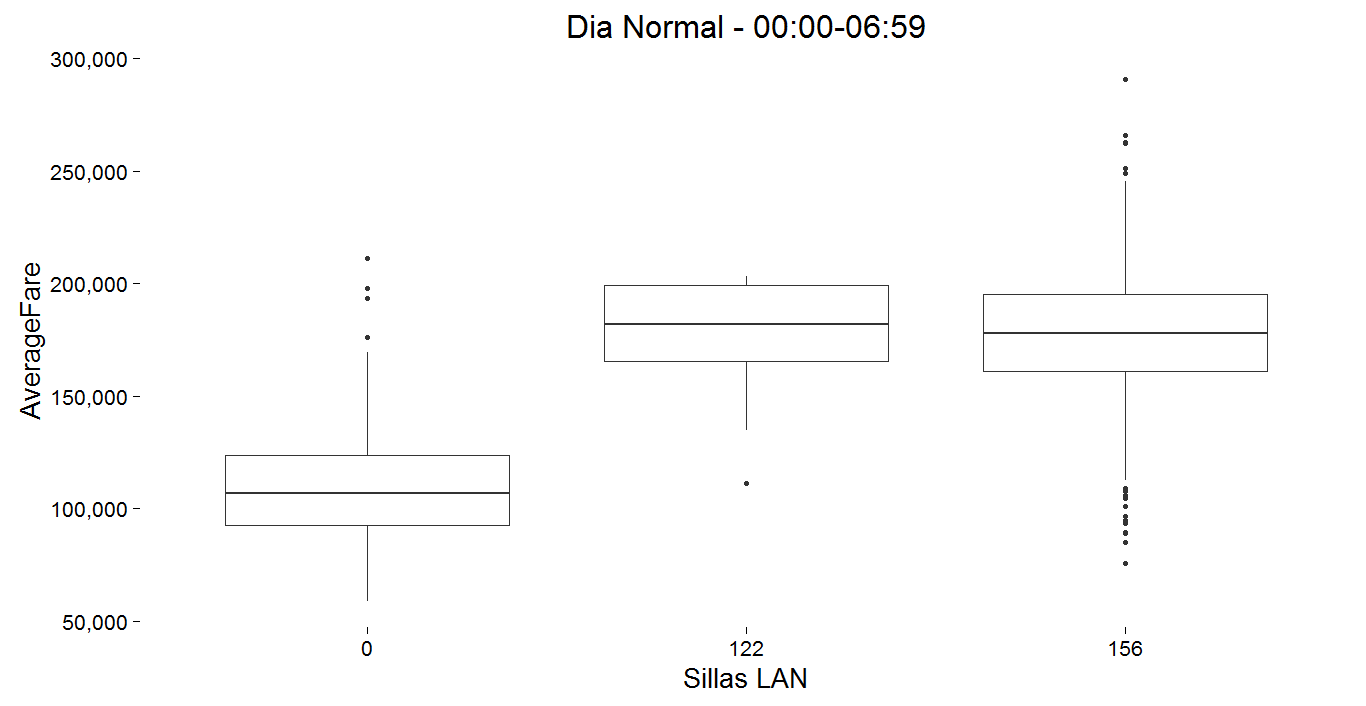
Como se visualiza en los diferentes gráficos a medida que disminuyen las sillas o se acerca la fecha de salida del vuelo el precio del tiquete es más caro. Estas dos variables actúan en conjunto para definir el precio del tiquete.

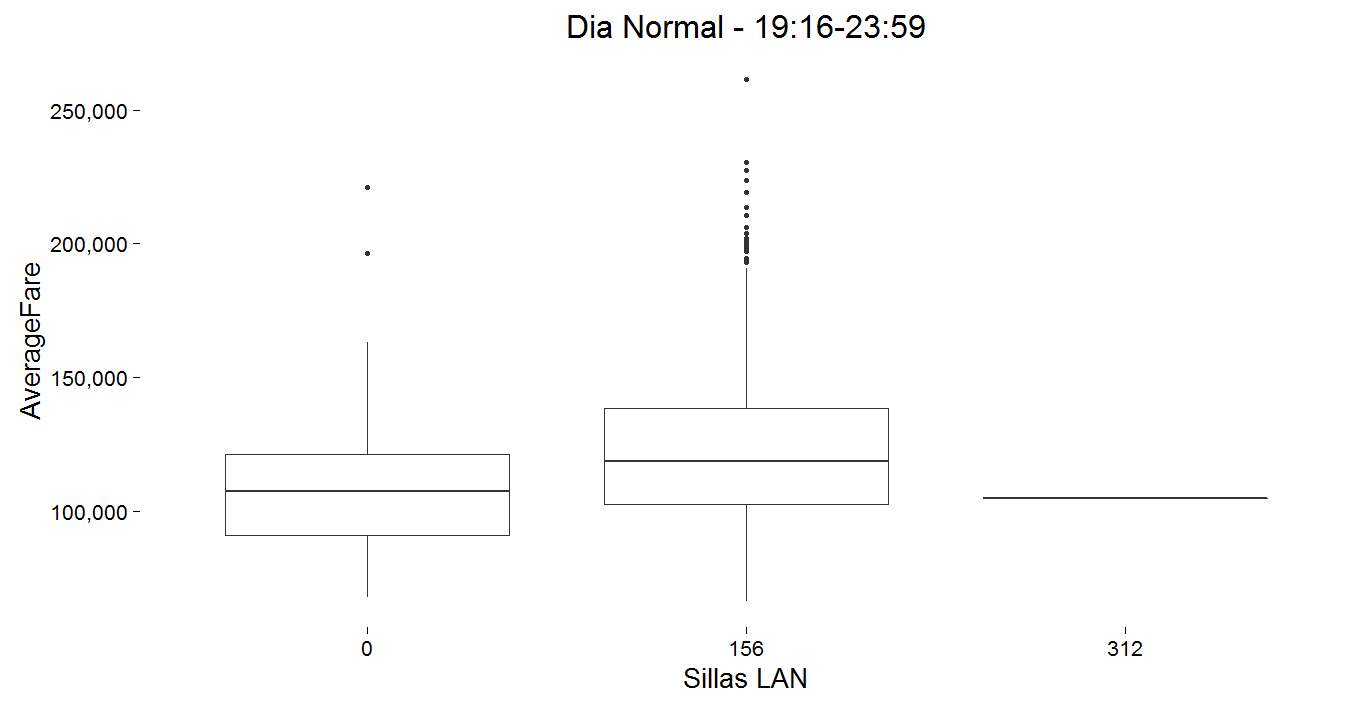
Se quería incluir informacion de la competencia relevante para BOGPSO pero resulto que no hay competencia, por lo que se evaluó en este punto informacion del OD Local BOGBGA. Se obtuvieron las sillas para cada fecha de vuelo y TOD, a parte se resumió la informacion de venta a este mismo nivel y se calculó para cada tipo de día si hubo impacto por la cantidad de sillas que la competencia ofrecía.



Se concluye que el impacto de la venta de sillas por parte de la competencia para este OD no es relevante y no influye mucho en la tarifa media. Por el contrario, en la mayoría de tipo de días al existir competencia la tarifa media aumenta.

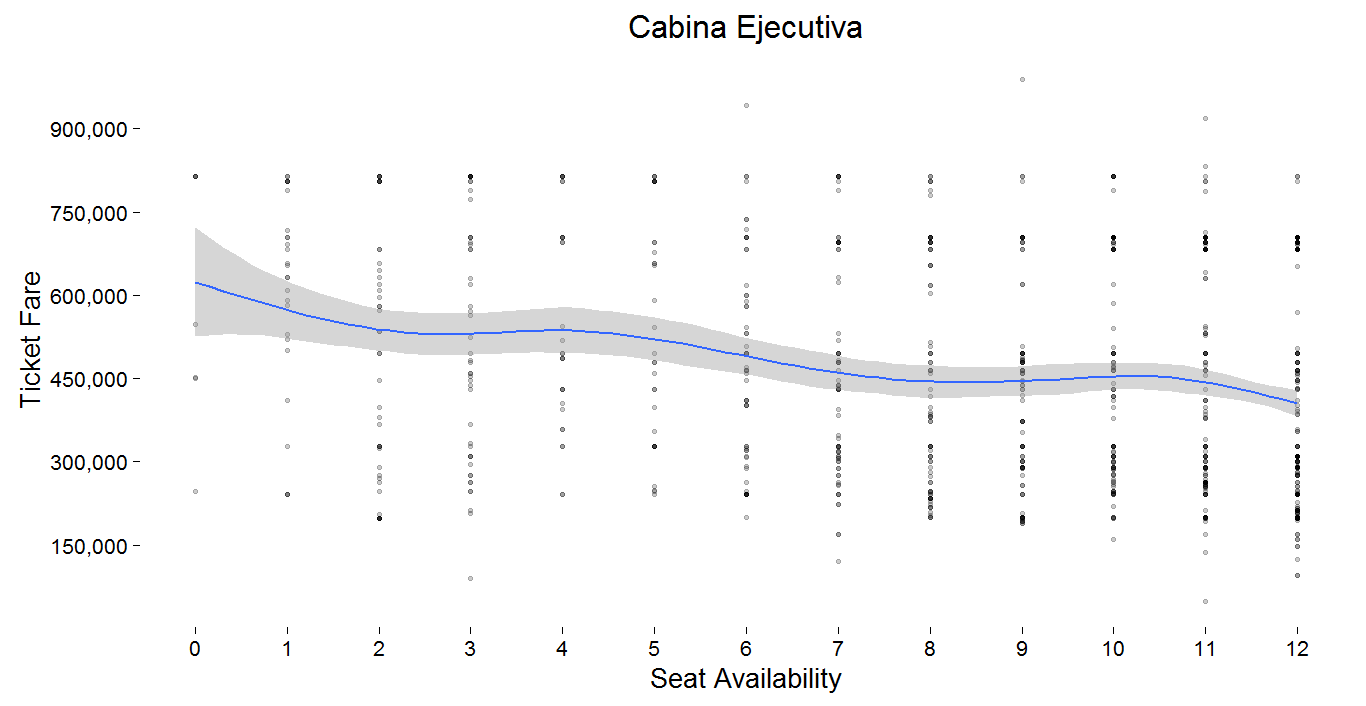
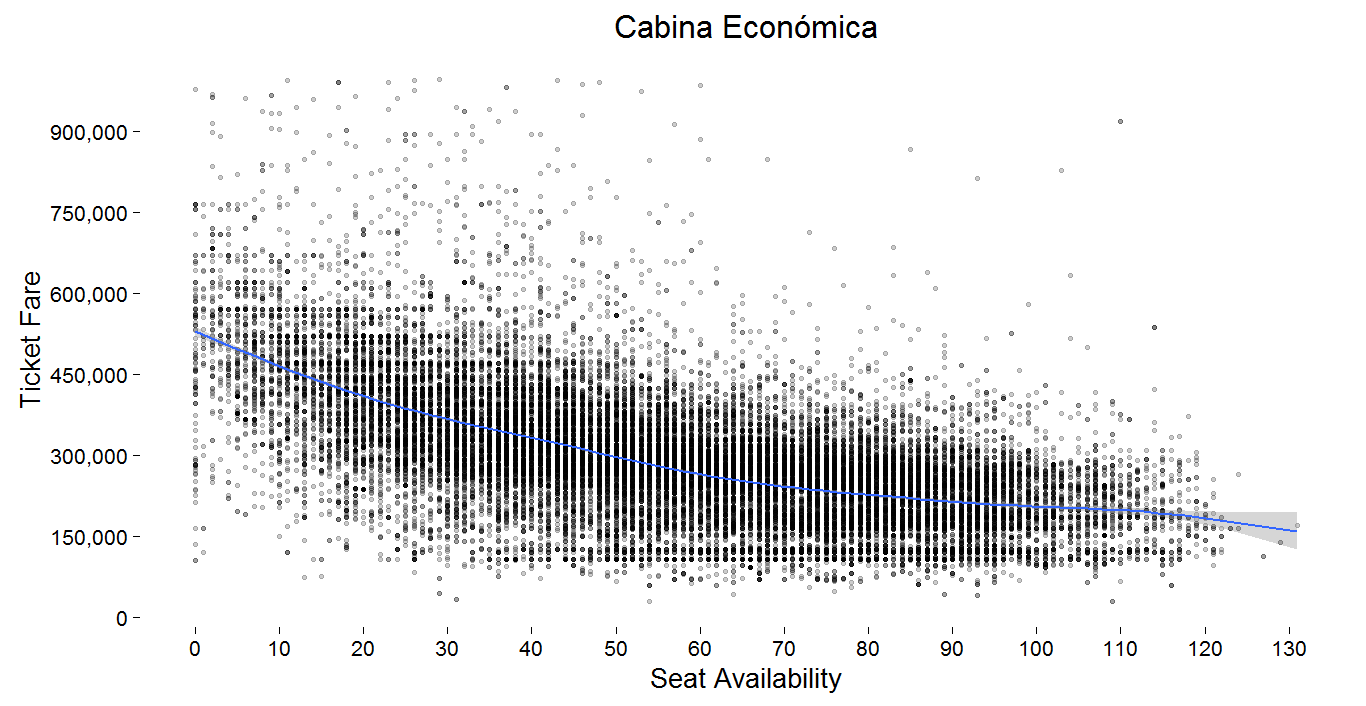
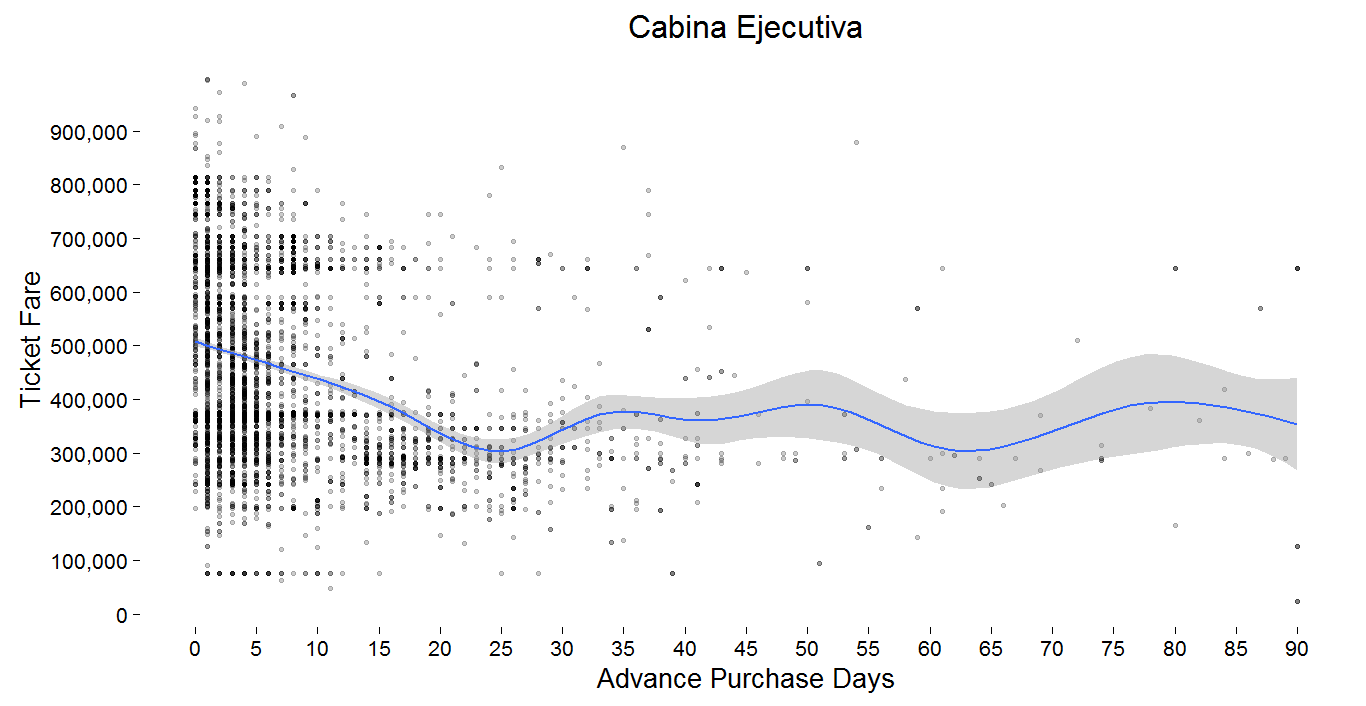
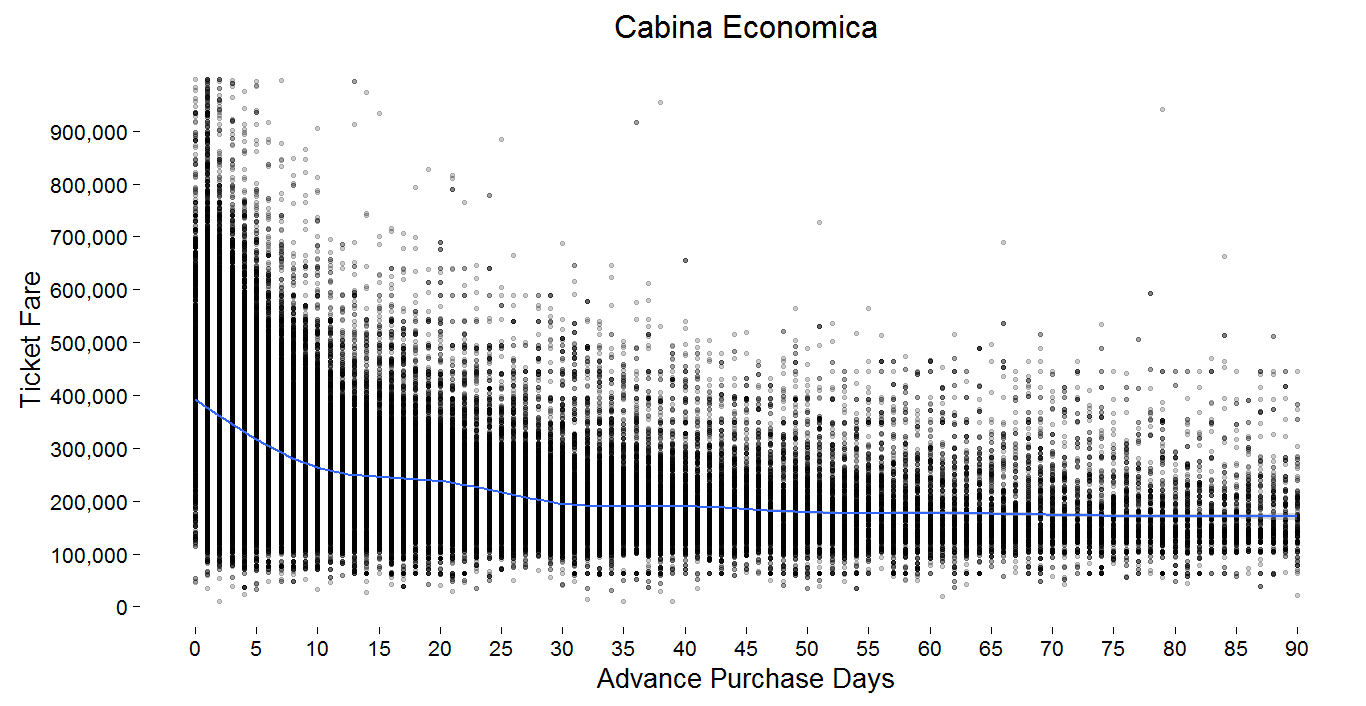
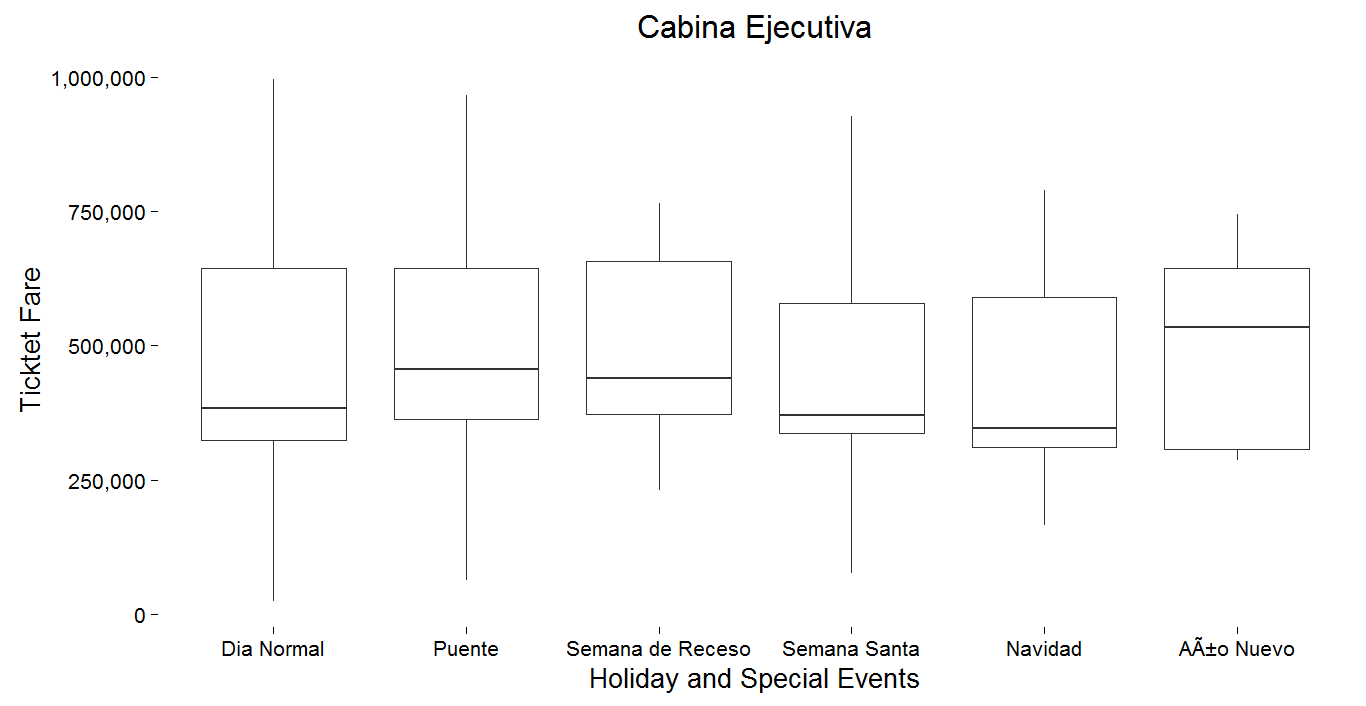
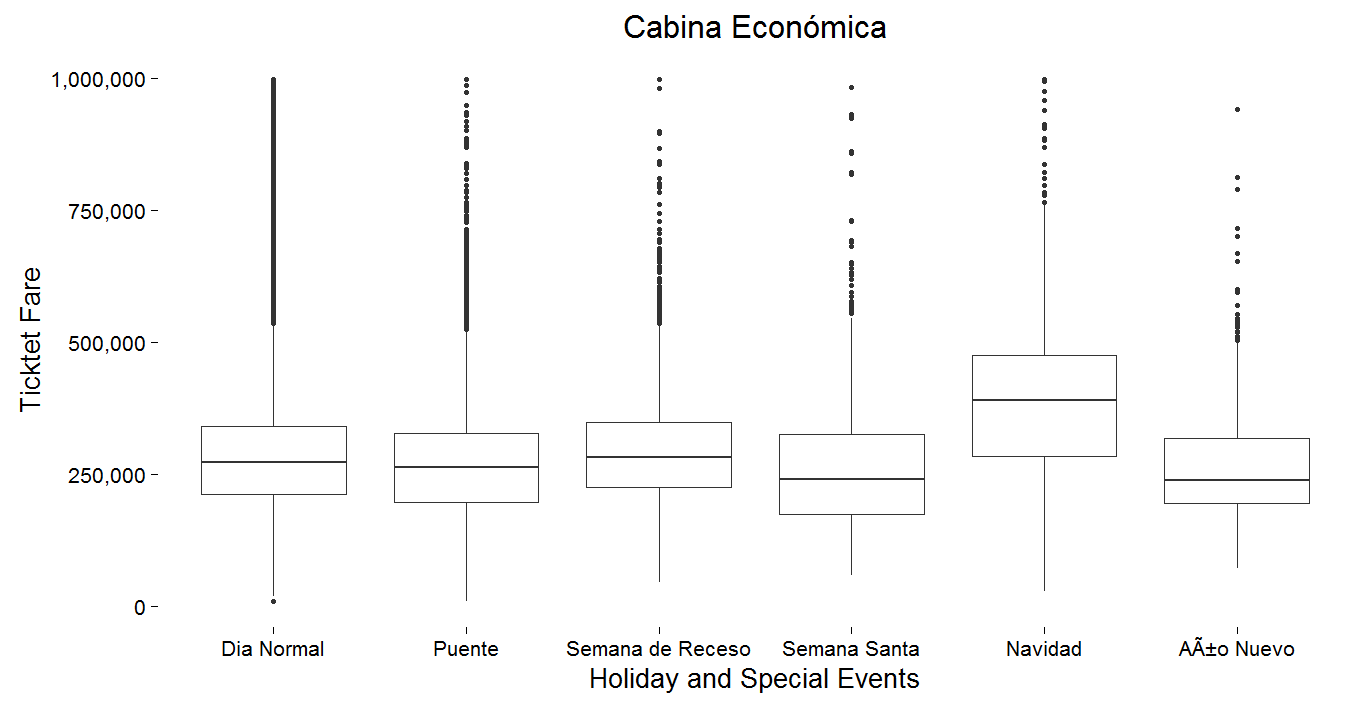
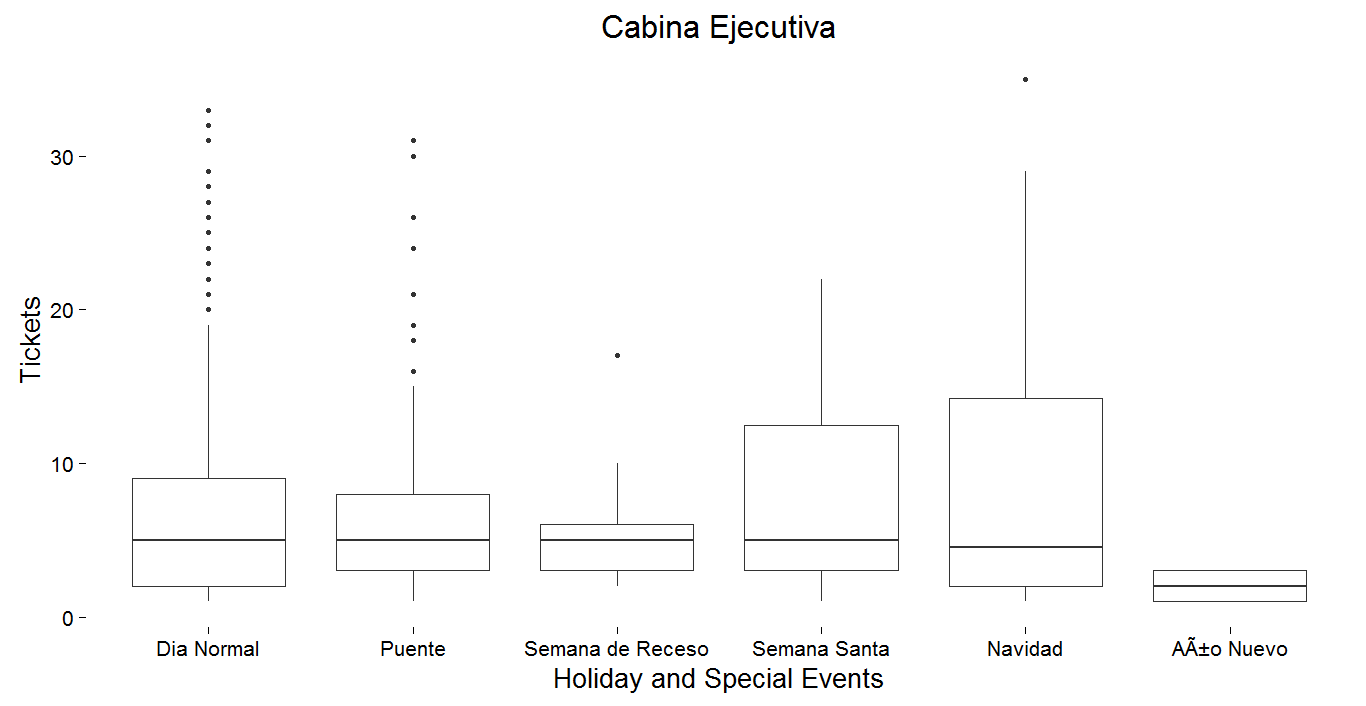
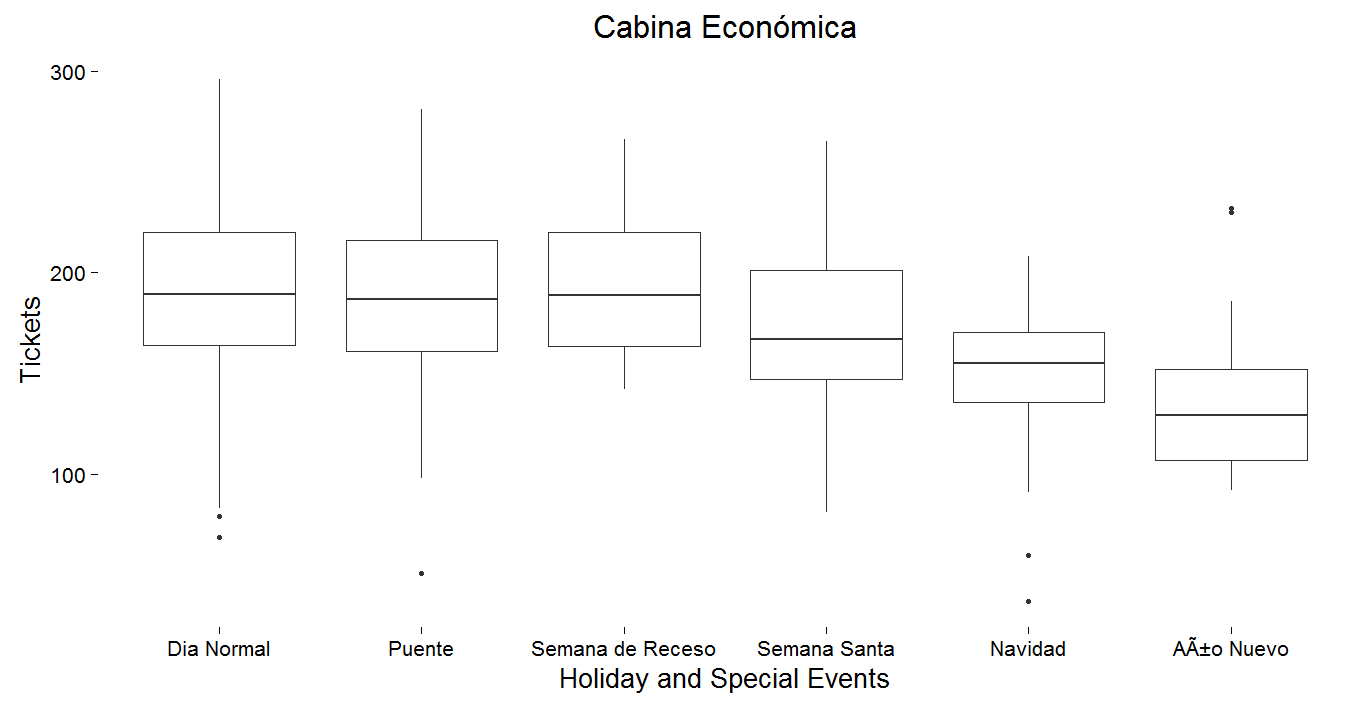
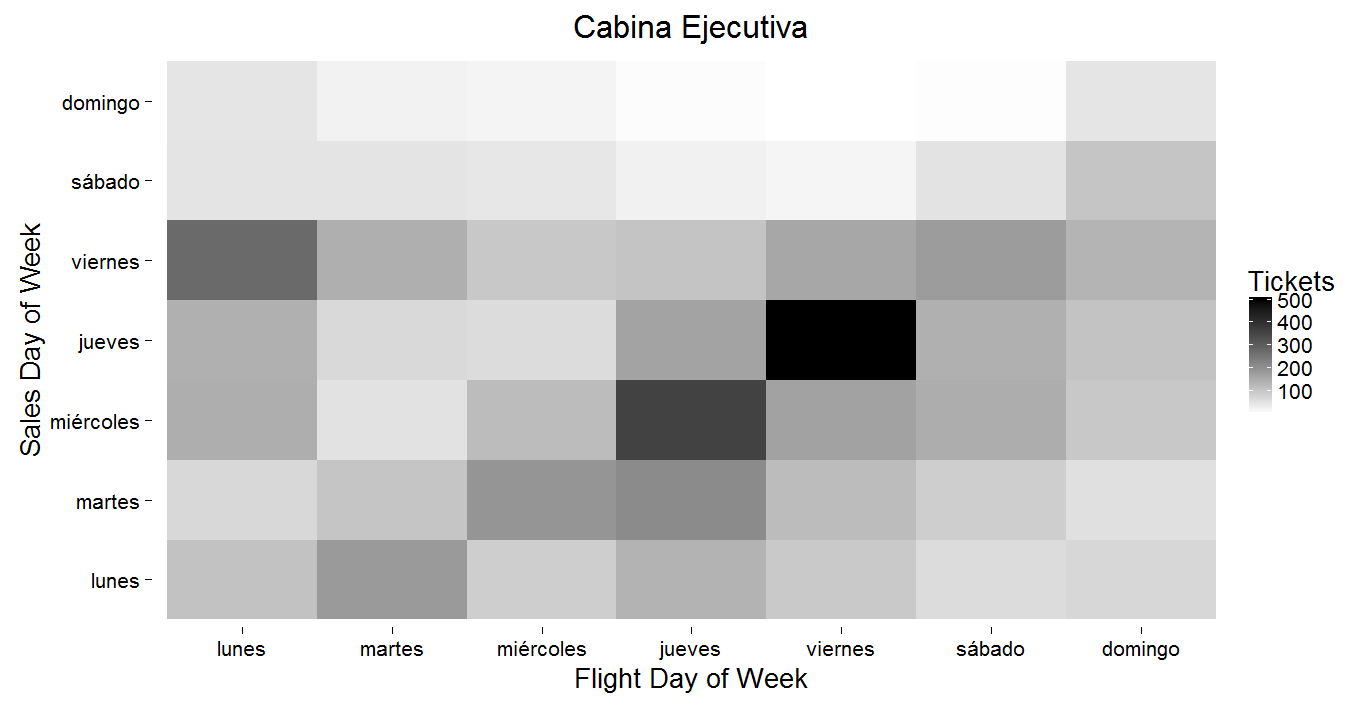
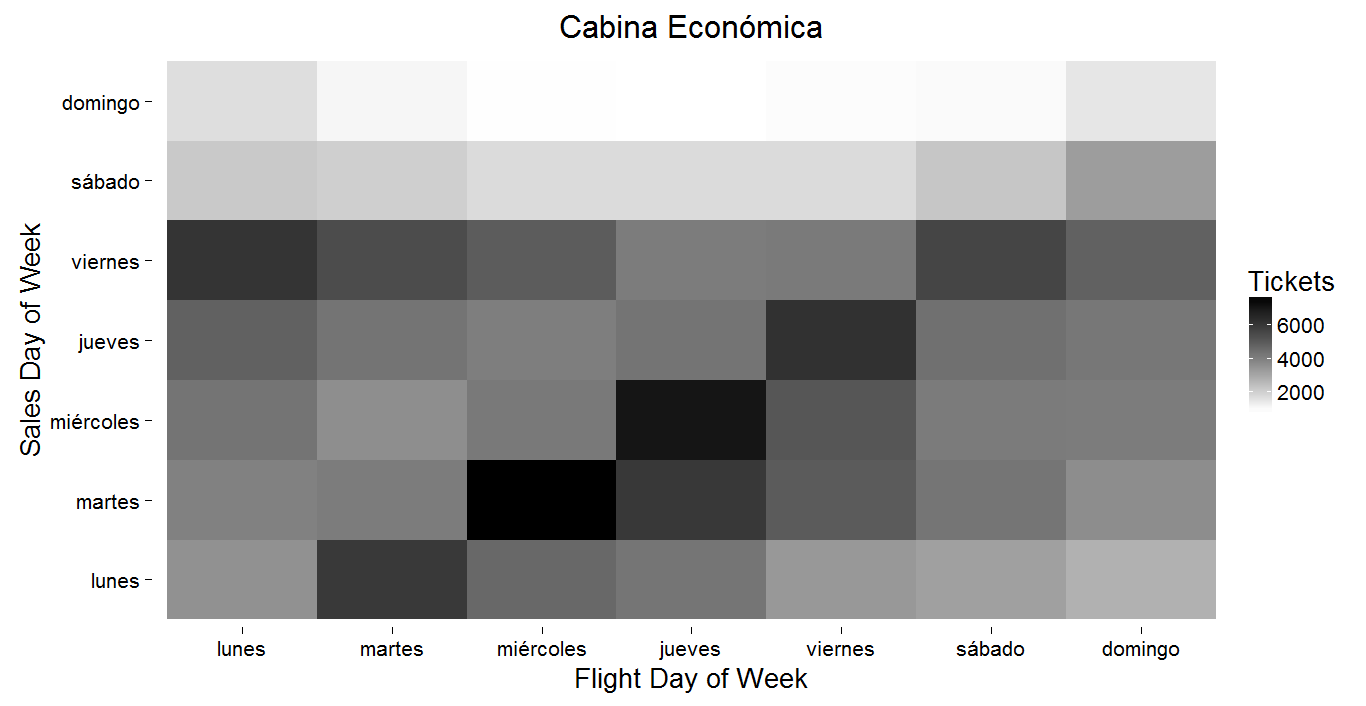
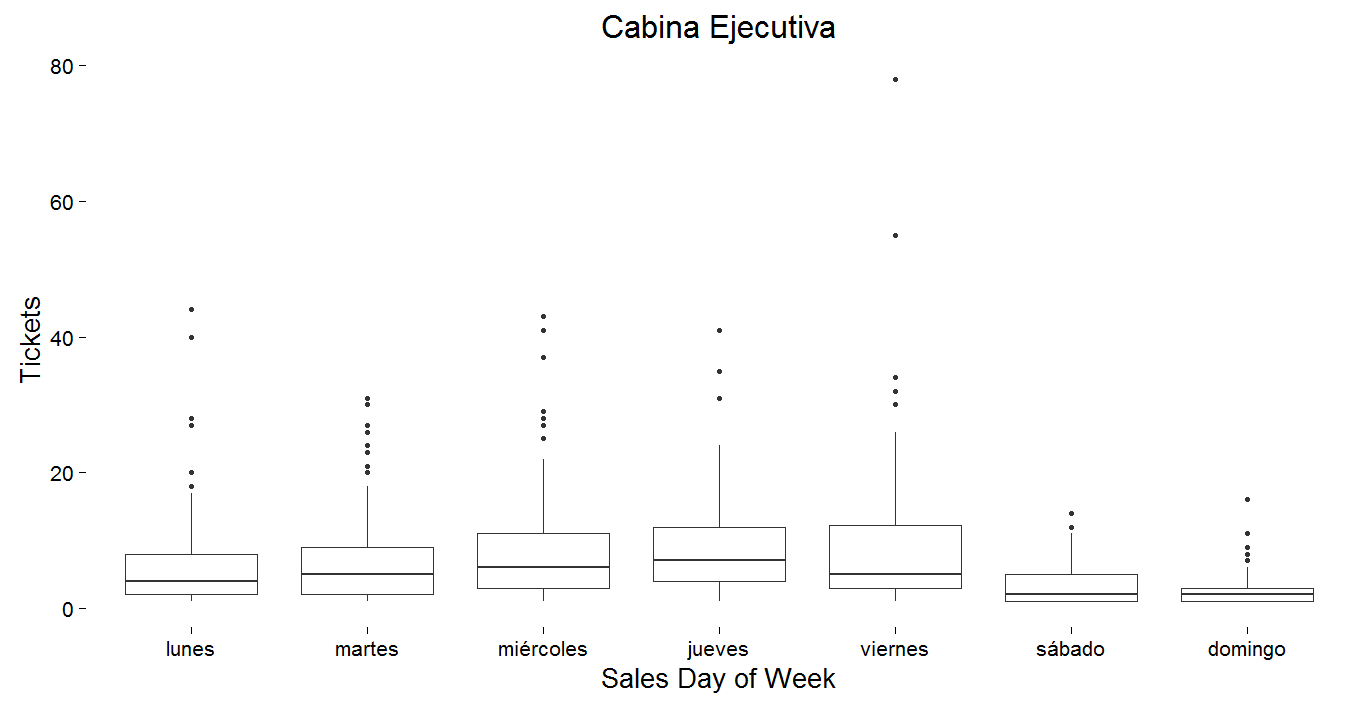
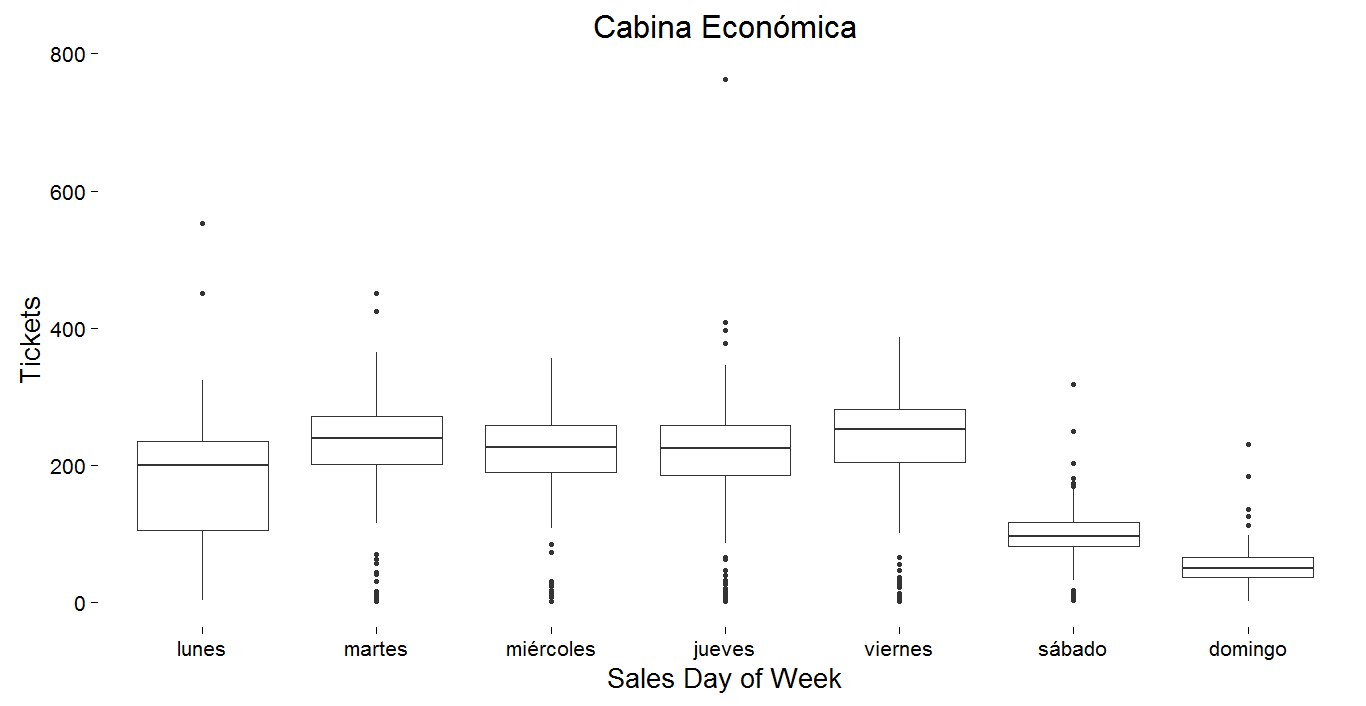
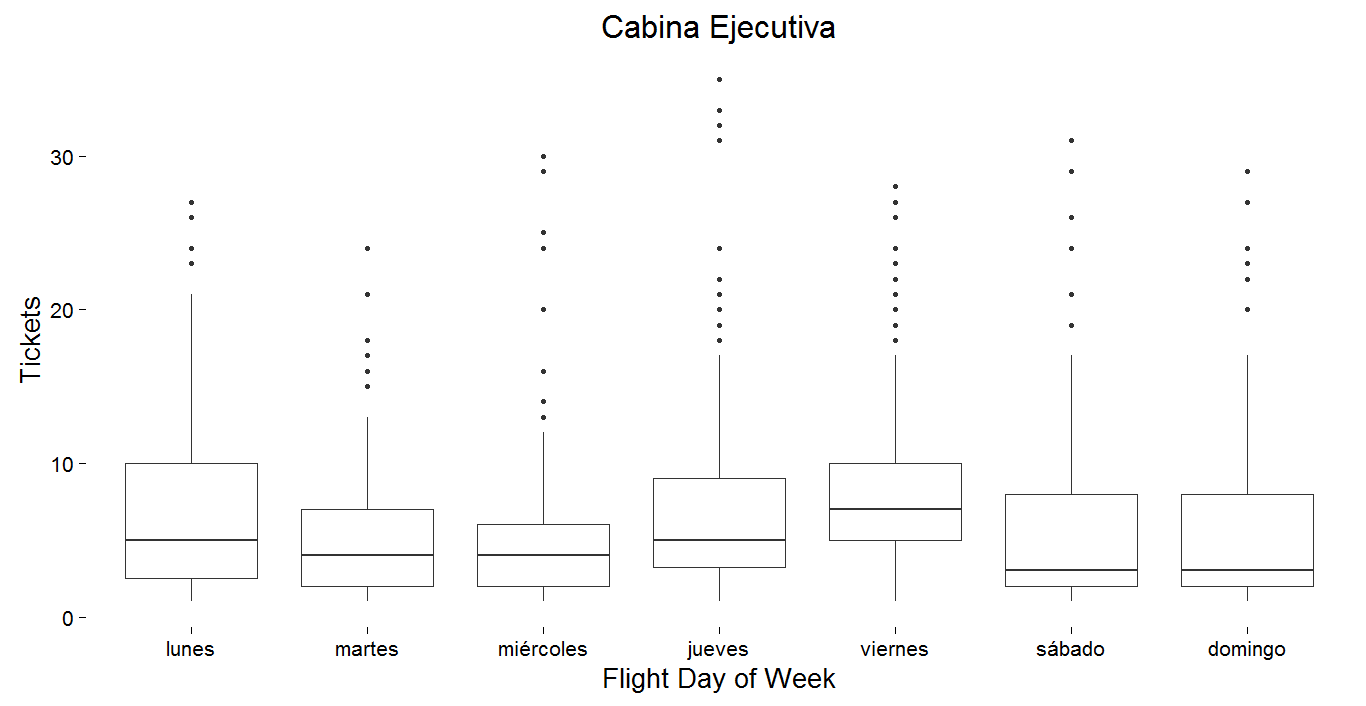
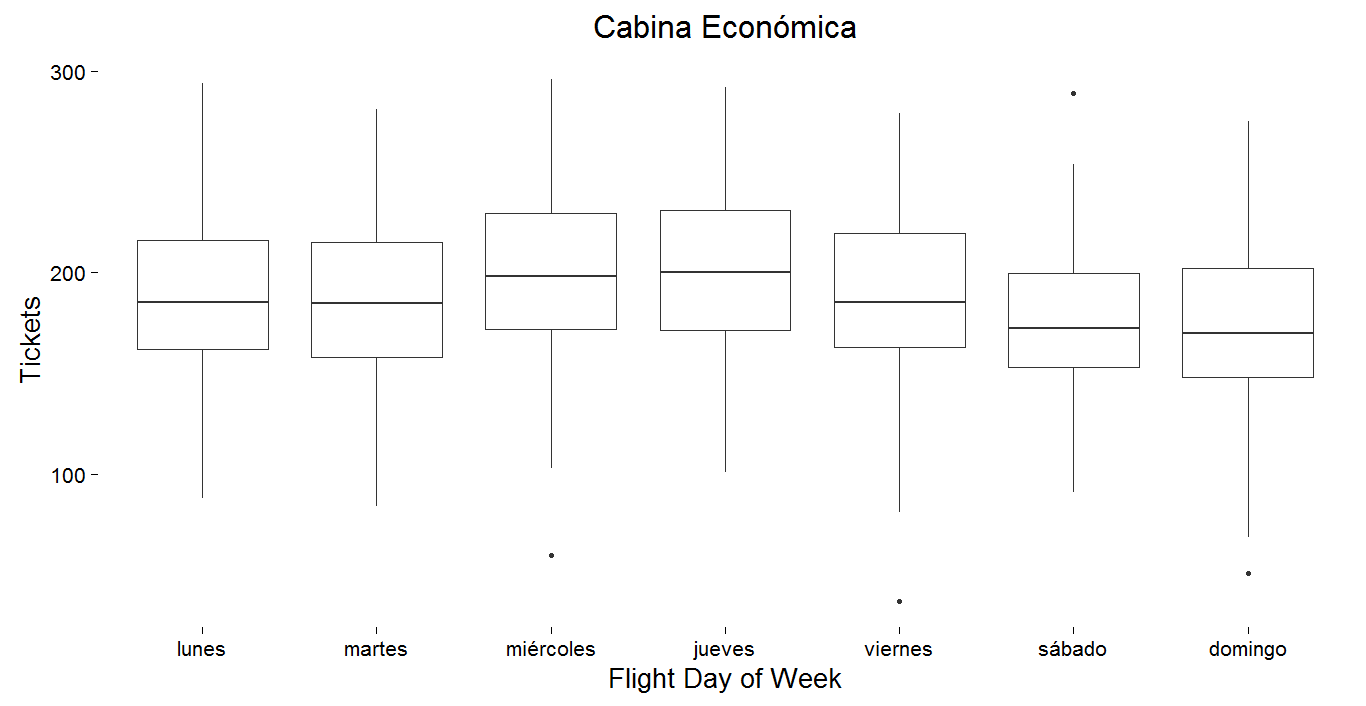
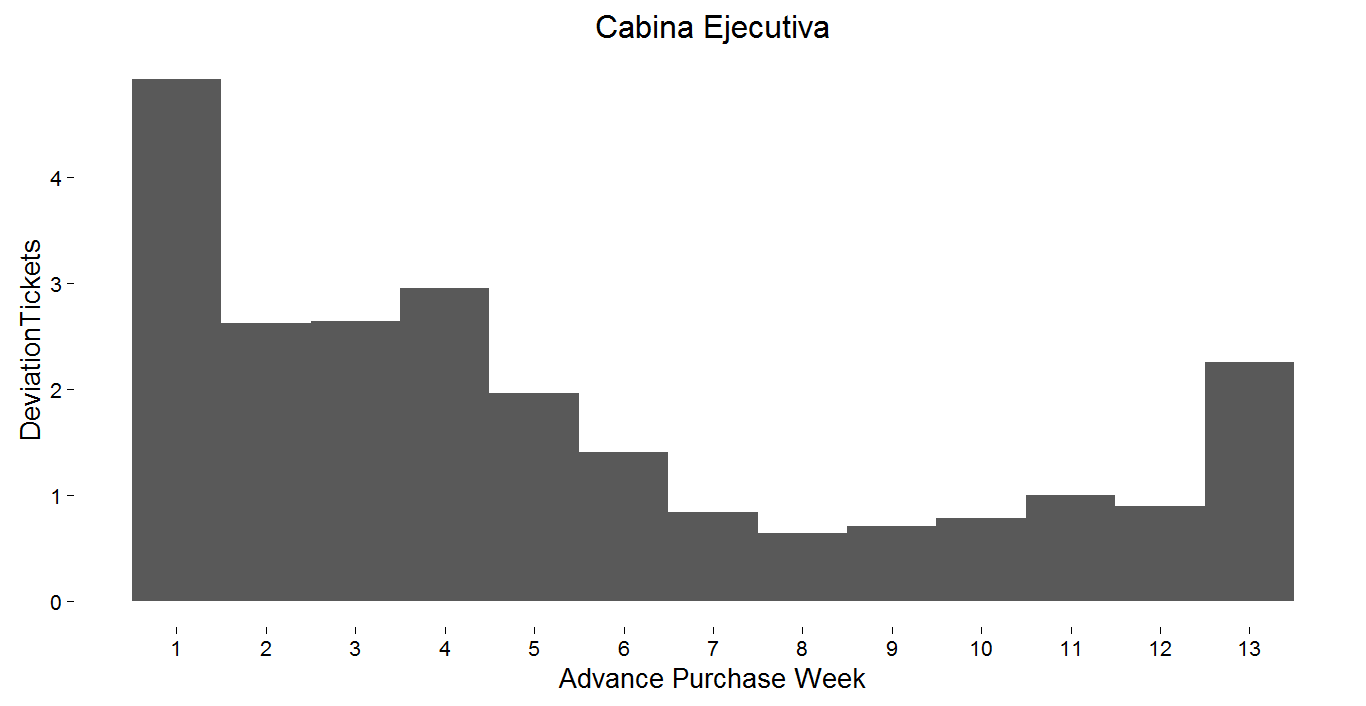
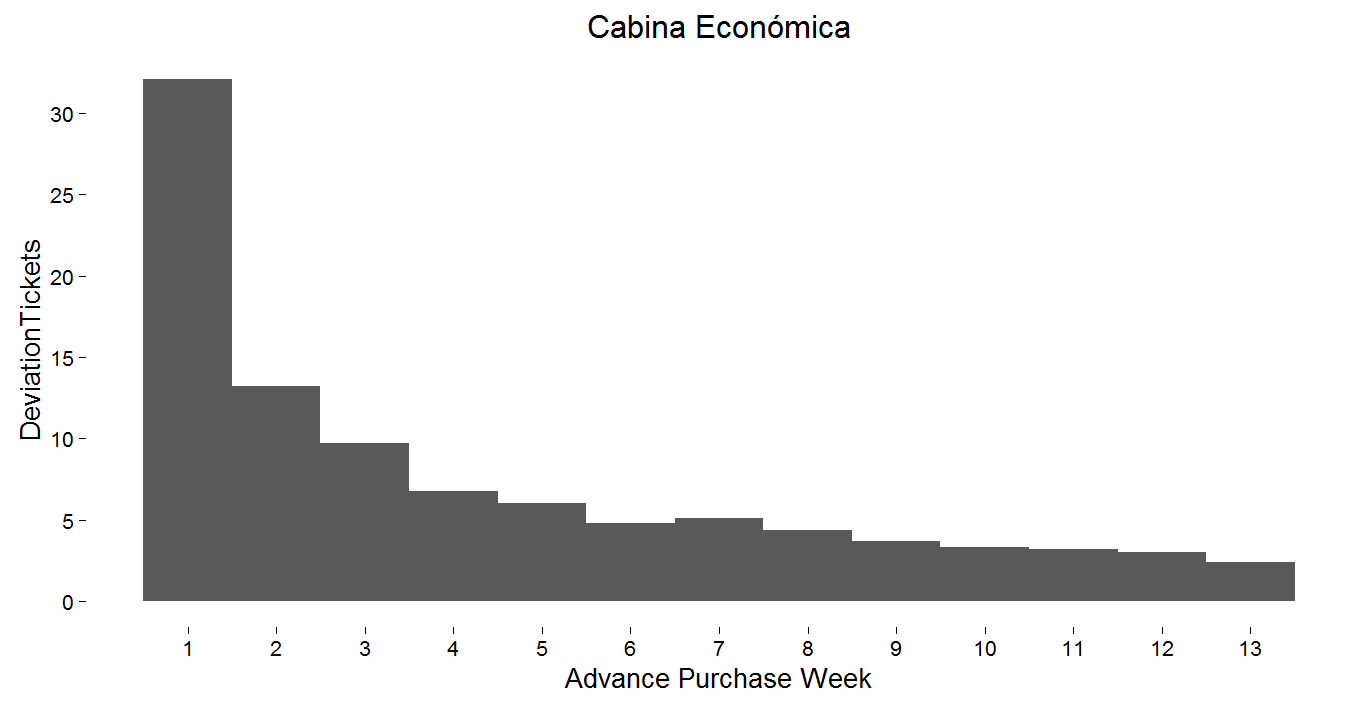
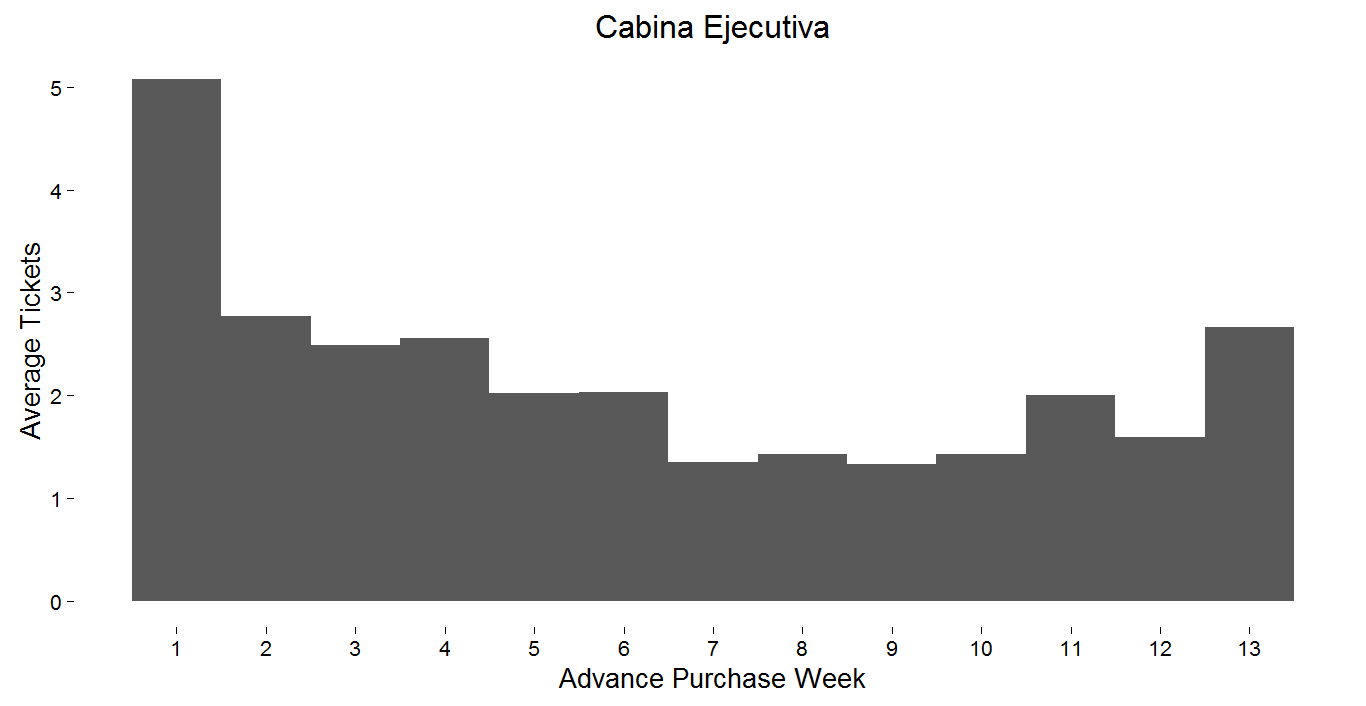
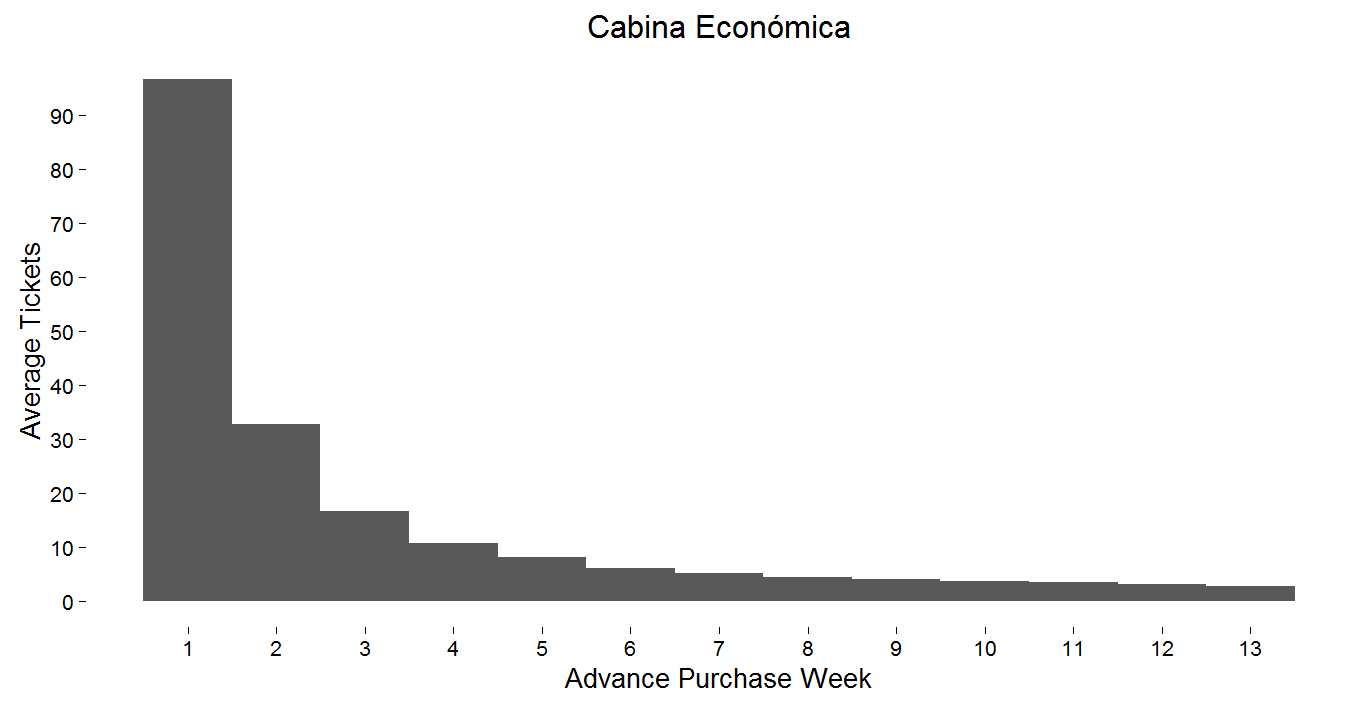
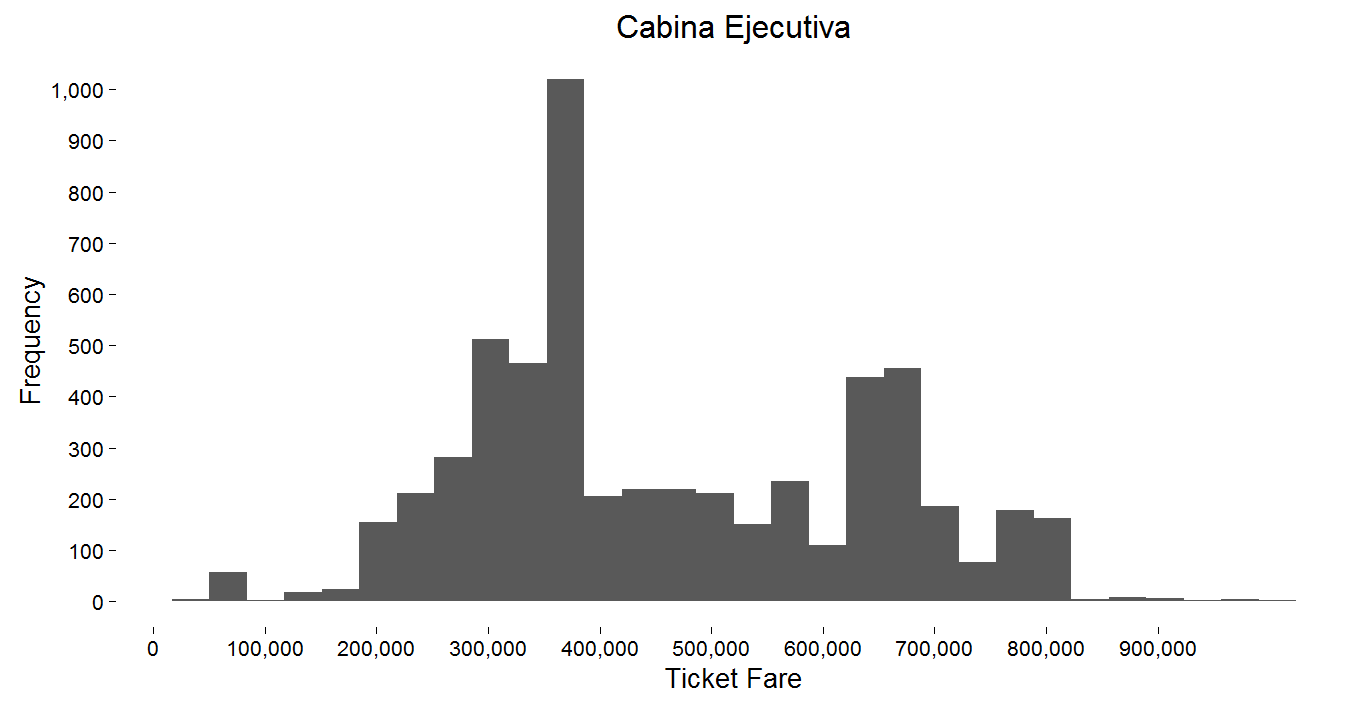
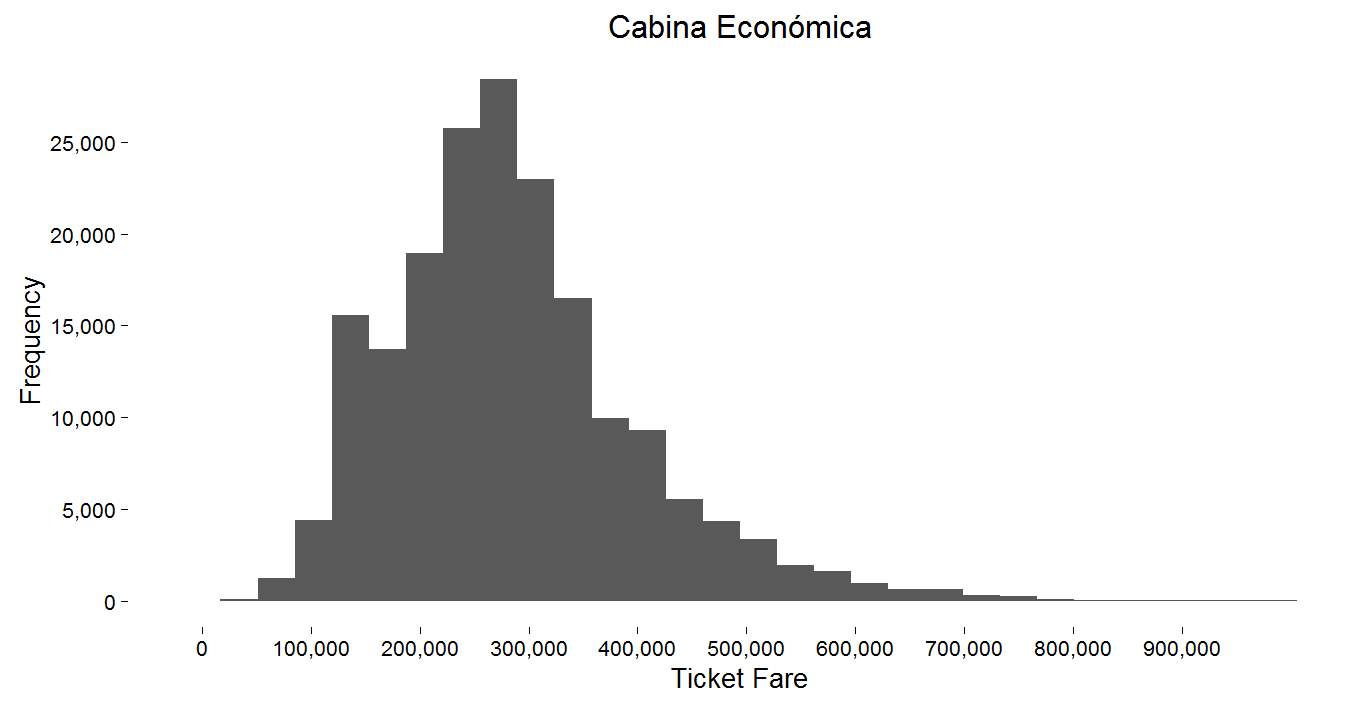
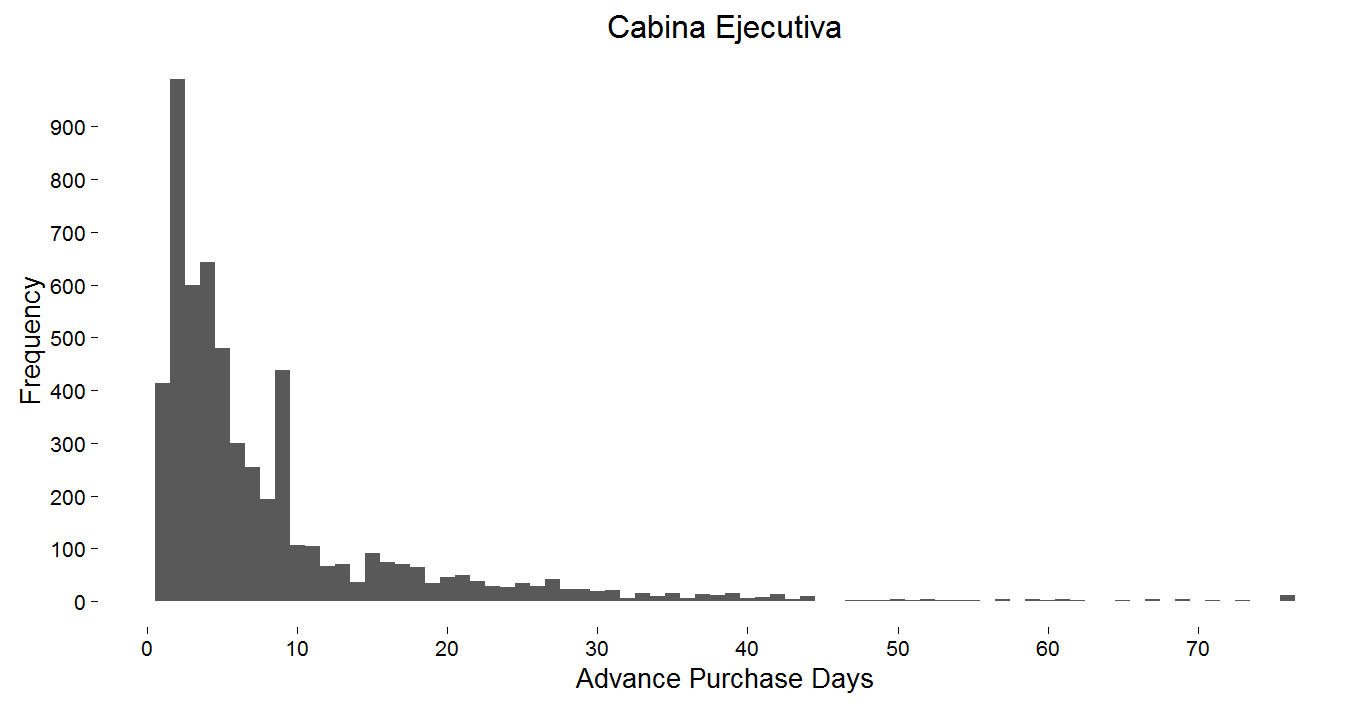
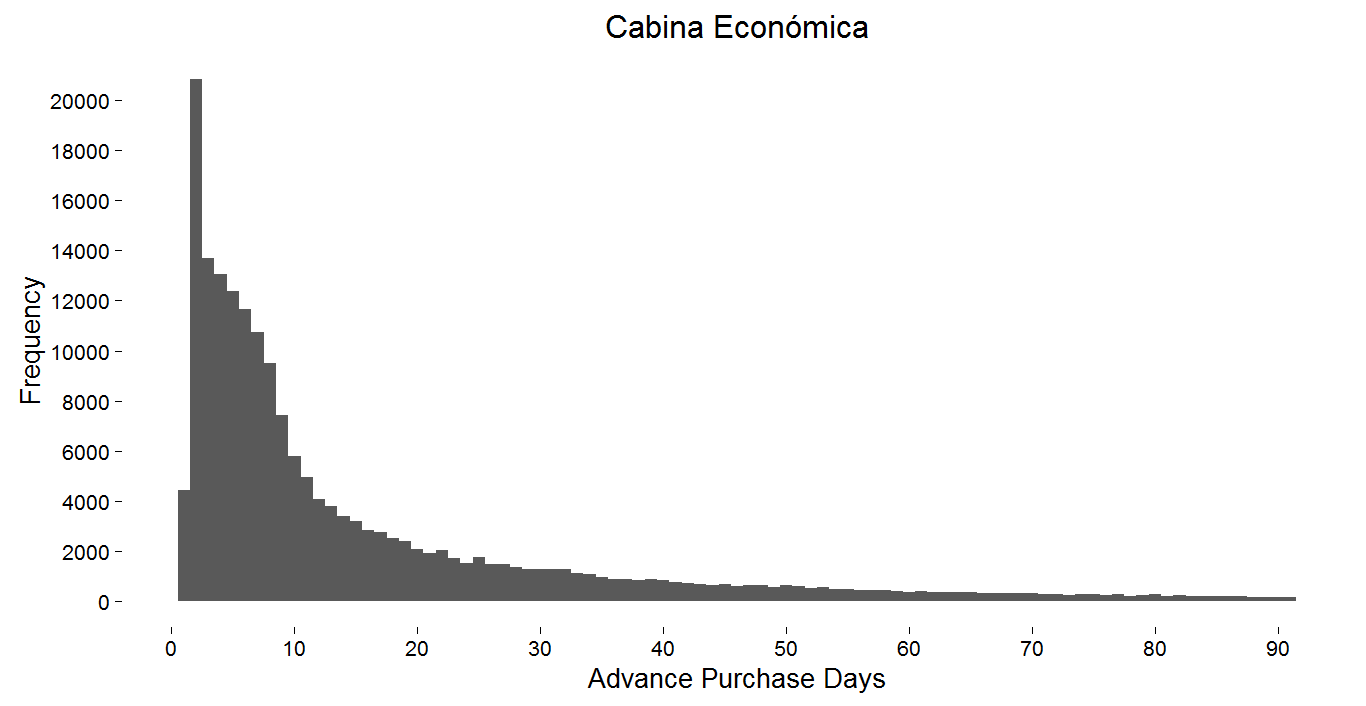
También se analiza por TOD para el tipo de día “Día Normal”





Se concluye que hay variaciones para cada TOD en el tipo de día “Día Normal” cuando existe competencia, no necesariamente tiene un impacto negativo en la tarifa. En algunos casos se observa que la tarifa mejora cuando hay competencia y en otras si disminuye la tarifa media.

Finalmente, se grafican las variables tenidas en cuenta anteriormente pero se hace una división para los datos de cada cabina comprada (económica y ejecutiva).



A partir de las gráficas anteriores, se concluye que es necesario analizar la informacion de cada una de las cabinas por separado ya que cuentan con comportamientos diferentes.

## Data Preparation

La fase de preparación de la información involucra aquellas actividades que se requieren para construir el conjunto de datos que serán usados en la parte de modelación en el proyecto. Estas actividades incluyen selección, limpieza, construcción e integración de la información.

### Select Data

El propósito de esta sub etapa es decidir sobre la información que será utilizada en el análisis, donde se seleccionan el nivel de agregación y los atributos de la información.

#### Aggregation Level

Para modelar la respuesta del comportamiento del mercado (demanda) existen dos énfasis de análisis. Por un lado, está el análisis a nivel individual, por otro lado, está el análisis a nivel agregado. En el caso del análisis a nivel individual se trata de modelar el proceso de un individuo a partir de la estimación de la función de utilidad de cada individuo. Mientras que los modelos agregados consisten en modelar las respuestas del comportamiento agregado de los clientes y así poder reflejar de manera apropiada las relaciones entre las variables.

En el caso de la investigación actual, se propone un análisis a nivel agregado. Este nivel no ha sido definido, pero tiene que tener en cuenta varias restricciones que el área de demanda sugirió para que las influencias en el sistema se puedan realizar de manera masiva y apropiadamente.

Como primera medida, se requirió que el análisis como mínimo se debe tener el análisis a nivel Departure Time Window con agrupación de días de anticipación de compra. Se ha visto en la literatura como … que un factor importante es la anticipación de compra por día, por lo que no se agruparan los días de anticipación de compra para realizar el análisis, sino que el pronóstico después de la modelación se agrupara al nivel requerido por el área, para que posteriormente se realice la respectiva optimización.

En principio, se agrega la información a nivel Departure Time Window definidos por PROS para los OD locales.

#### Features

A partir del nivel de agregación que se requiere se modelaron los los atributos se creen puedan afectar la demanda en un momento dado. Tomando como base la informacion de Venta Integrada, se obtienen los siguientes atributos: 1. Fecha de vuelo, 2. Fecha de transaccion, 3. OD Local, 4. Numero de Vuelo, 5. Numero de tiquete, 6. Cabina comprada, 7. Pasajeros, 8. Fare amount (USD) y 9. Fare amount (COP). A partir de la fecha de vuelo y la fecha de transaccion (fecha de compra) se calcula 10. Anticipacion de compra. Despues con la informacion de VQs se obtienen para cada una de los registros 10. Numero de vuelo y 11. Sillas disponibles realizando el cruce entrefecha de vuelo, DCP con Anticipacion de compra, OD Local y Leg. Por otro lado, de acuerdo al archivo provisto por el area de demanda donde se identifican para cada OD los 12. Holidays and Special Events, se identifican para cada uno de los registros el tipo de dia de acuerdo a la fecha de vuelo. De acuerdo a la hora de salida de los vuelos se identifica el 13. Departure Time Window en base a los rango de horas que actualemte tiene PROS.

Como se menciono anteriormente, para poder utilizar el modelo de Two Stage Least Squares es necesario poder identificar variables instrumentales que influencien el precio de los tiquetes mas no la demanda de Avianca. De acuerdo a la literatura dos de las variables mas adecuadas son las 14. sillas/frecuencias de la competencia y el 15. precio promedio para los demas destinos que tengan la misma longitud de recorrido. Es por esto, que a partir de Industry Schedules se obtuvieron las sillas clasificadas para los mismos Departure Time Windows según PROS para cada una de las fechas y rutas; tambien se obutvo de la misma fuente Venta Integrada, los precios promedios para cada una de las fechas de vuelo con la misma anticipacion.

### Clean Data

Para poder realizar y obtener estimaciones adecuadas en los modelos es necesario poder tener informacion acertada. Por lo tanto, es necesario evaluar la informacion y definir si es necesario alguna tecnica de eliminacion de datos atipicos para poder limpiar la informacion.

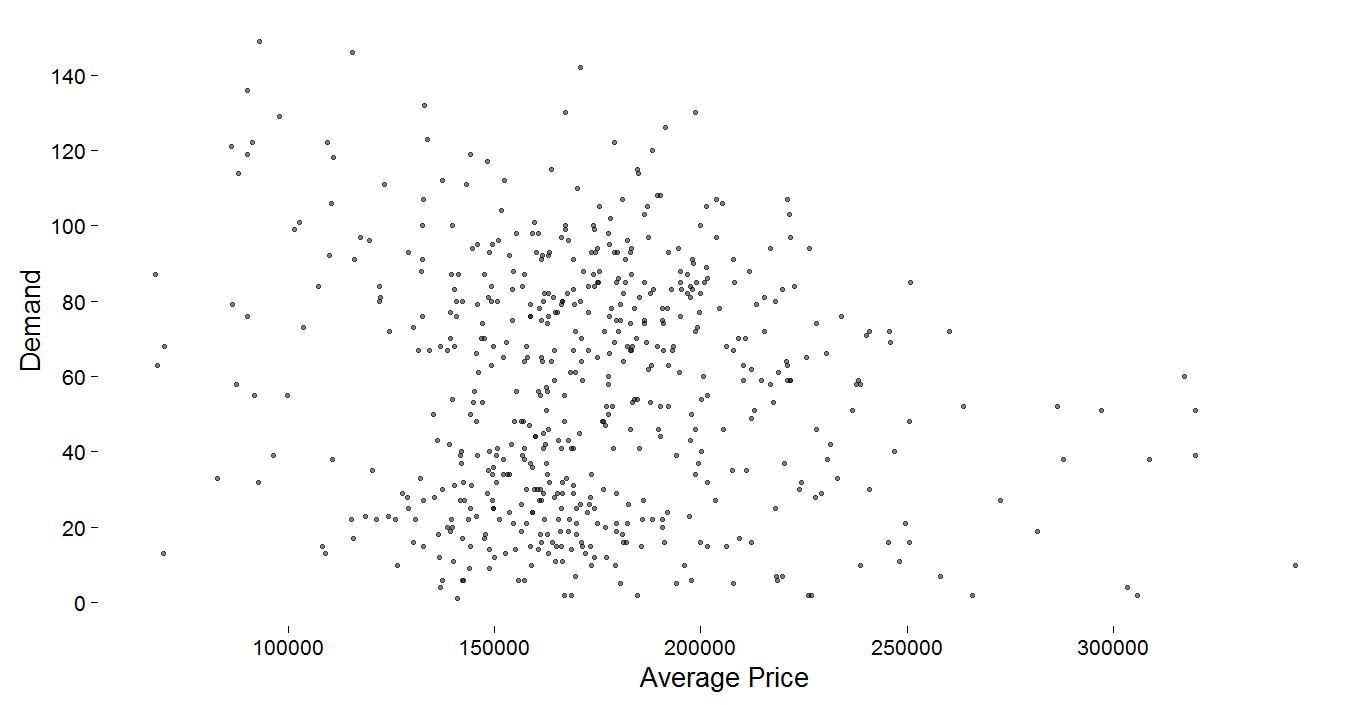
Se evaluaron diferentes tecnicas y las mas adecuadas para este caso fueron las siguientes:

* Mean Absolute Deviation
* Multivariate Guassian Distribution
* Adjusted Squared Plot

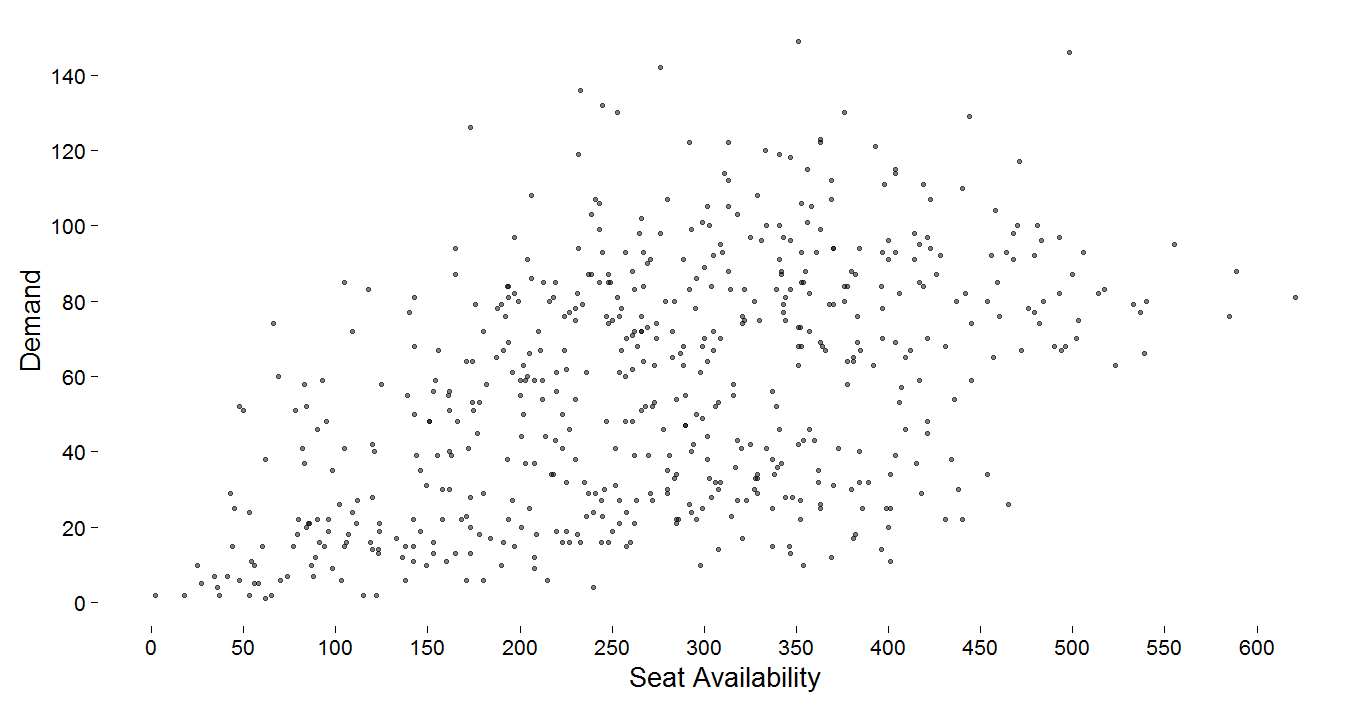
Es importante aclarar que todas estas tecnicas de eliminacion de valores atipicos asumen que las variables que se utilizan en cada una de las tecnicas tienen una distribucion uniforme.

Se aplicaron estas tecnicas para cada unos de los subconjuntos de informacion con los que se realizaran el modelamiento y para diferntes rutas y se obtuvieron los siguientes resultados :

**OD OW:** BOGMDE **CABINA:** Económica **TOD:** 08:30-11:59 **HSE:** Día Normal **AP:** 1

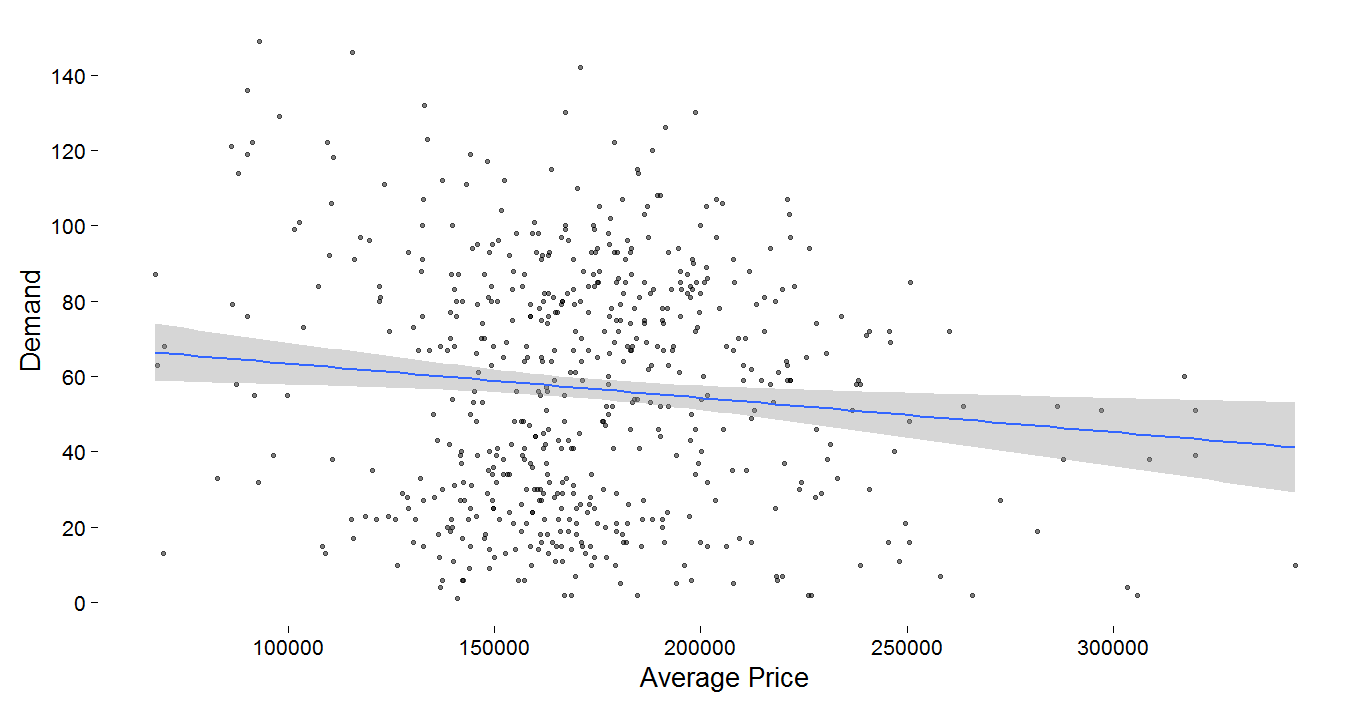


En el anterior grafico se identifican algunas zonas donde no es comun encontrar puntos, estas zonas son por ejemplo cuando la tarifa es baja y hay muy poca demanda o por lo contrario, cuando la tarifa es muy alta y la demanda tambien es alta.



Por otro lado, tambien es necesario corregir la informacion donde la demanda es mas alta que las sillas disponibles. Esto se puede generar debido a que la informacion de los VQs es una unica foto de la fecha de vuelo mientras que la demanda si es en tiempo real y las sillas pueden variar durante el dia.

De acuerdo a los datos anteriores, se grafica una regresion lineal con variable independiente Precio del Tiquete vs. Demanda como variable independiente.

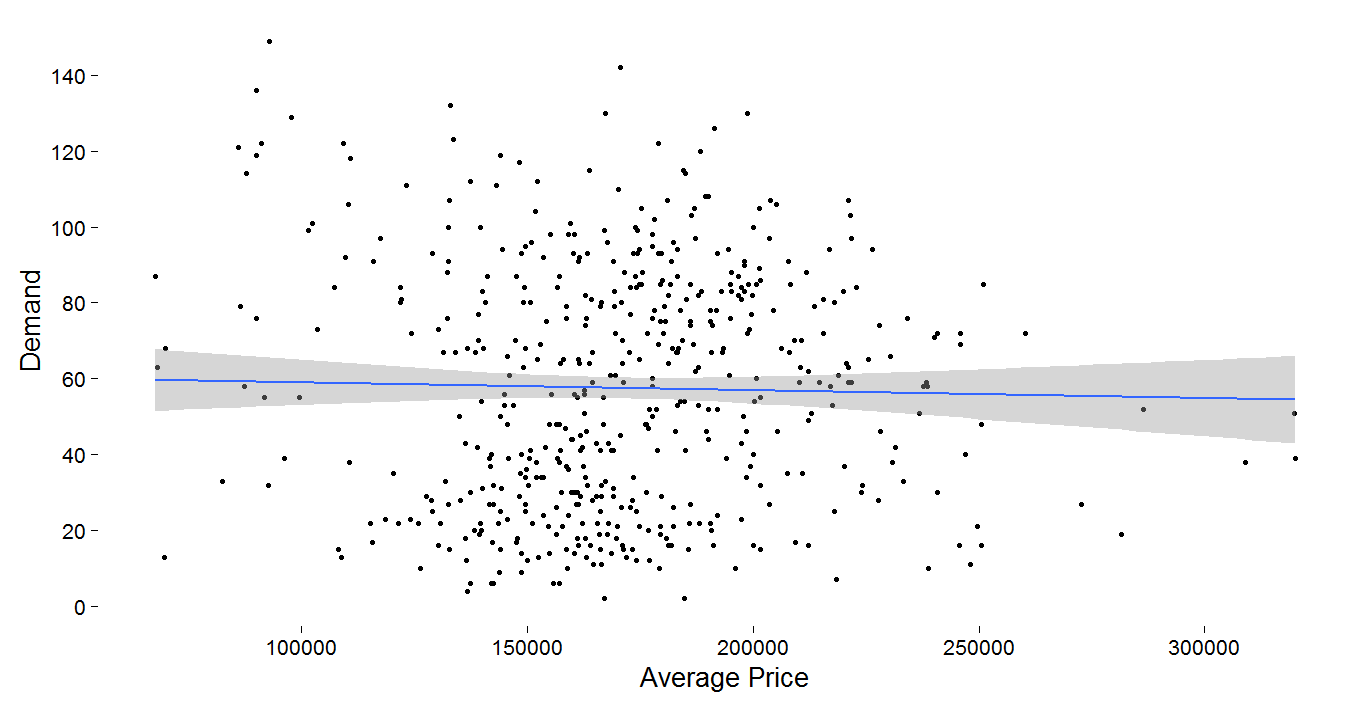
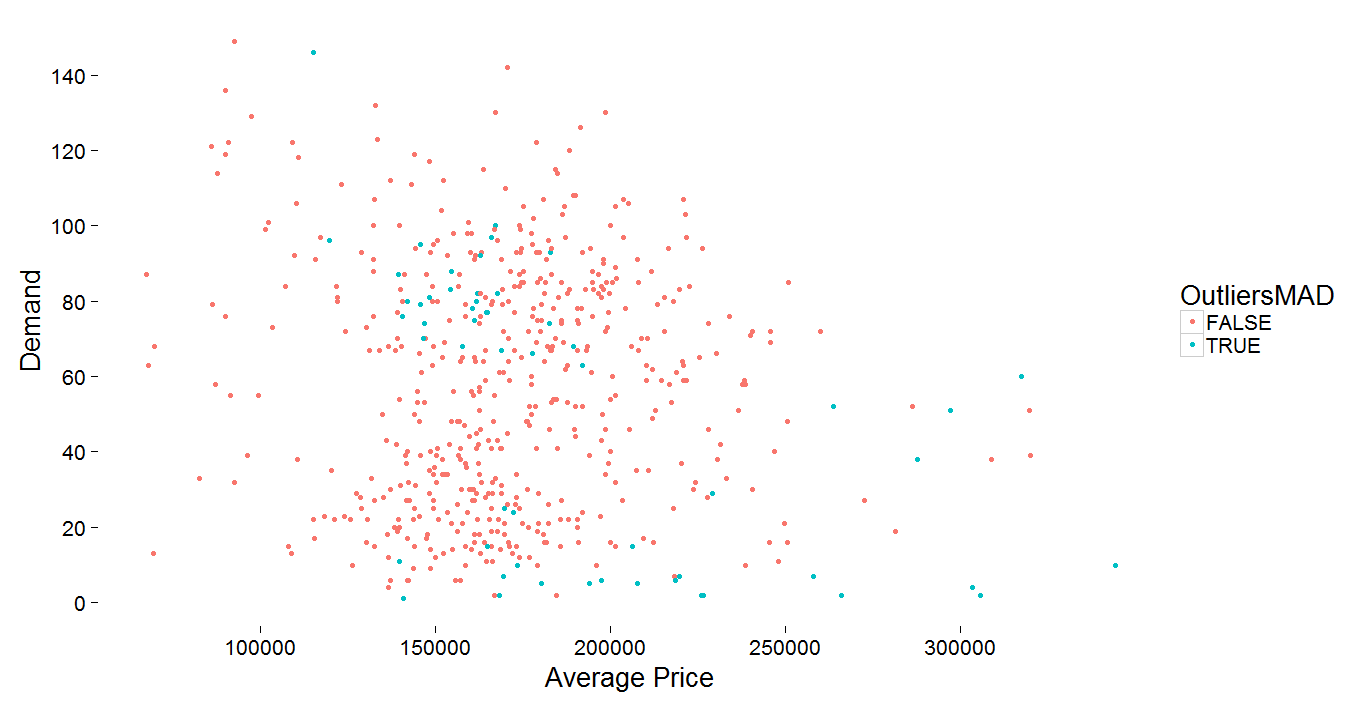
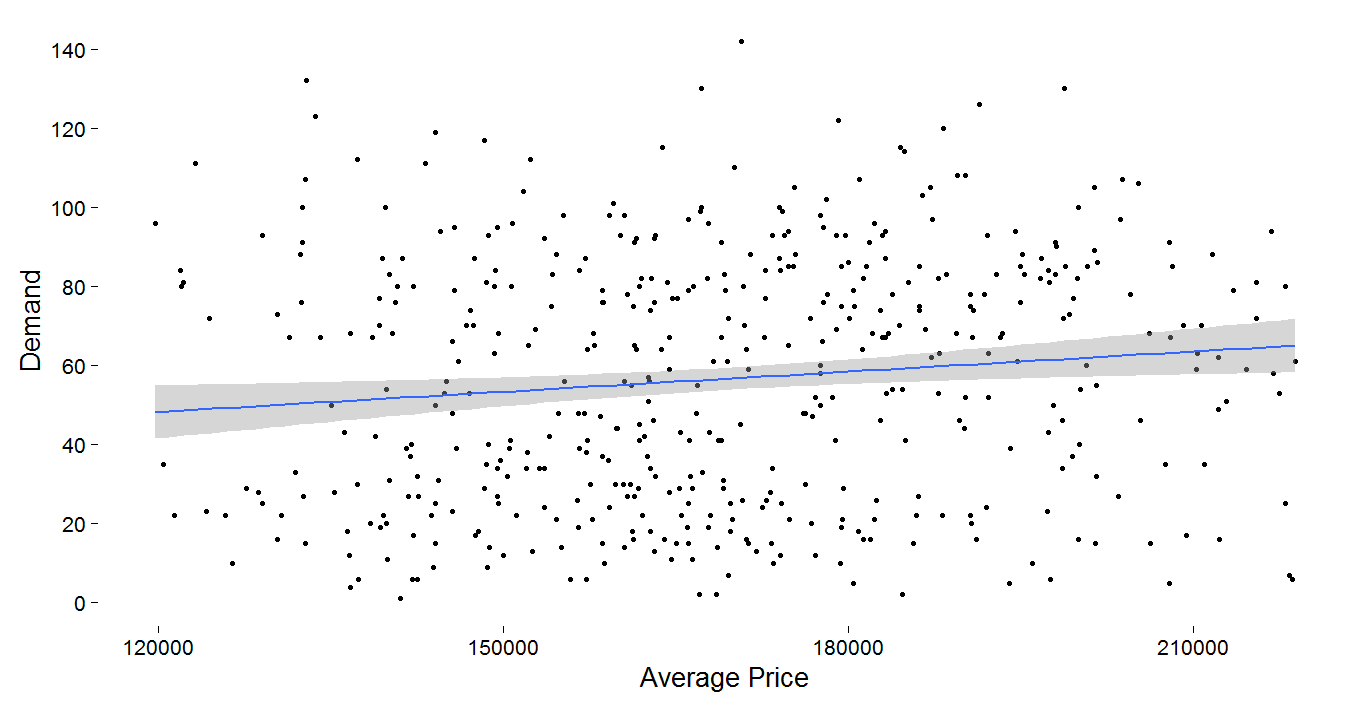
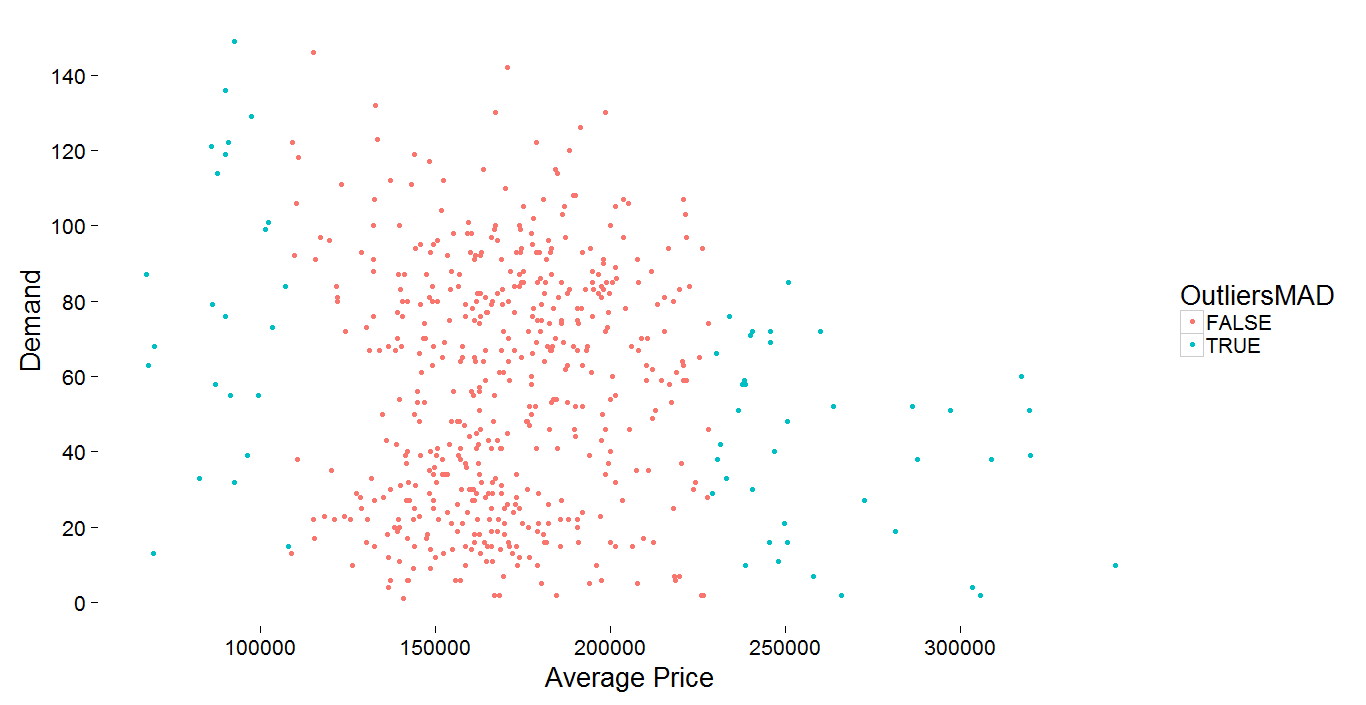


La regresion es adecuadas ya que se espera que a medida que el precio aumente menos demanda habra en el vuelo.

Se aplicaron los tres metodos de identificacion de valores antipicos

Mean Absolute Deviation

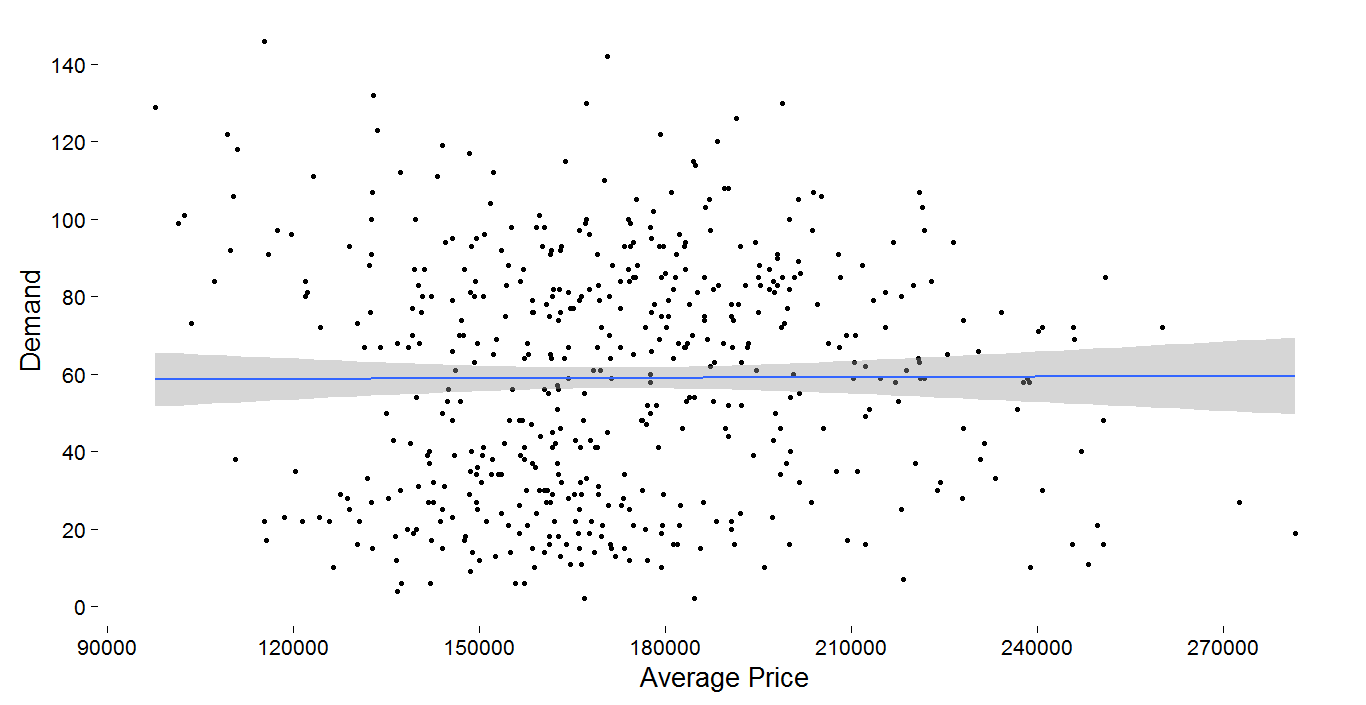
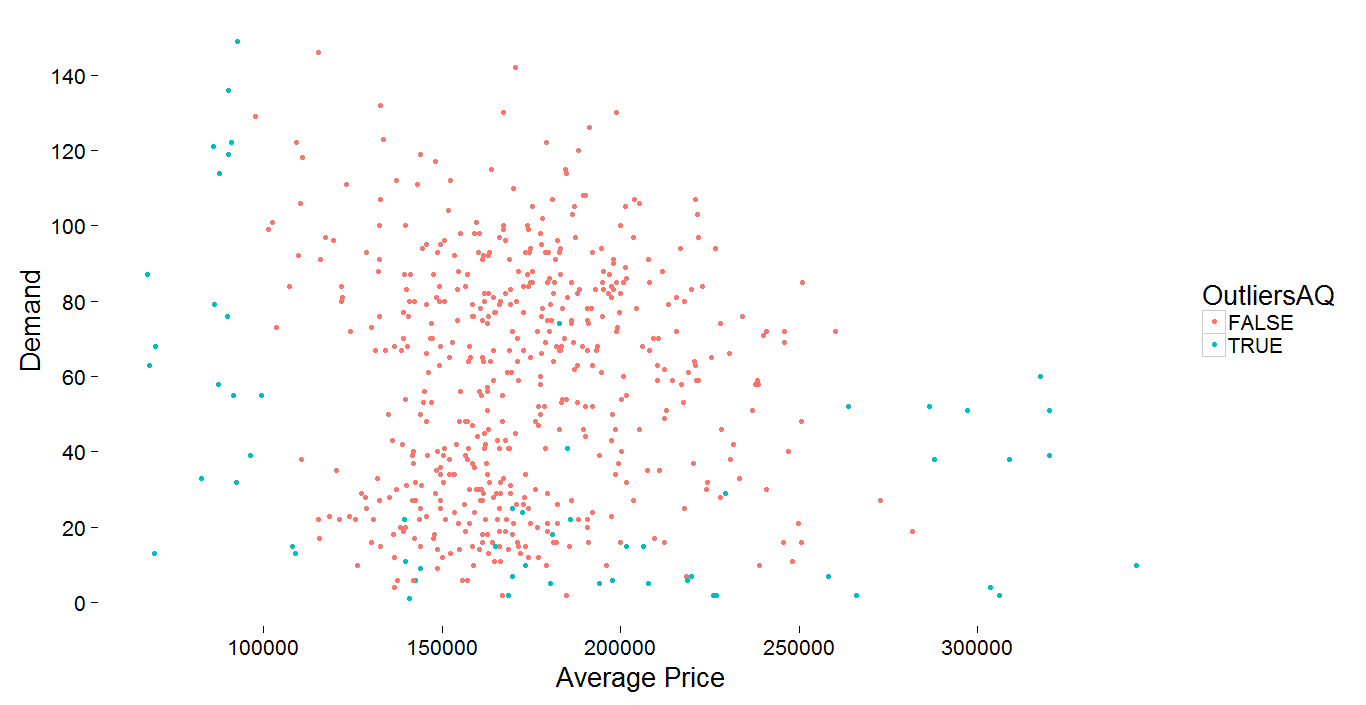
Este metodo se aplico por separado para las variables Average Price y Seat Avaiability, se evaluo el impacto de la eliminacion de los valores atipicos en la regresion lineal Precio vs Demanda.

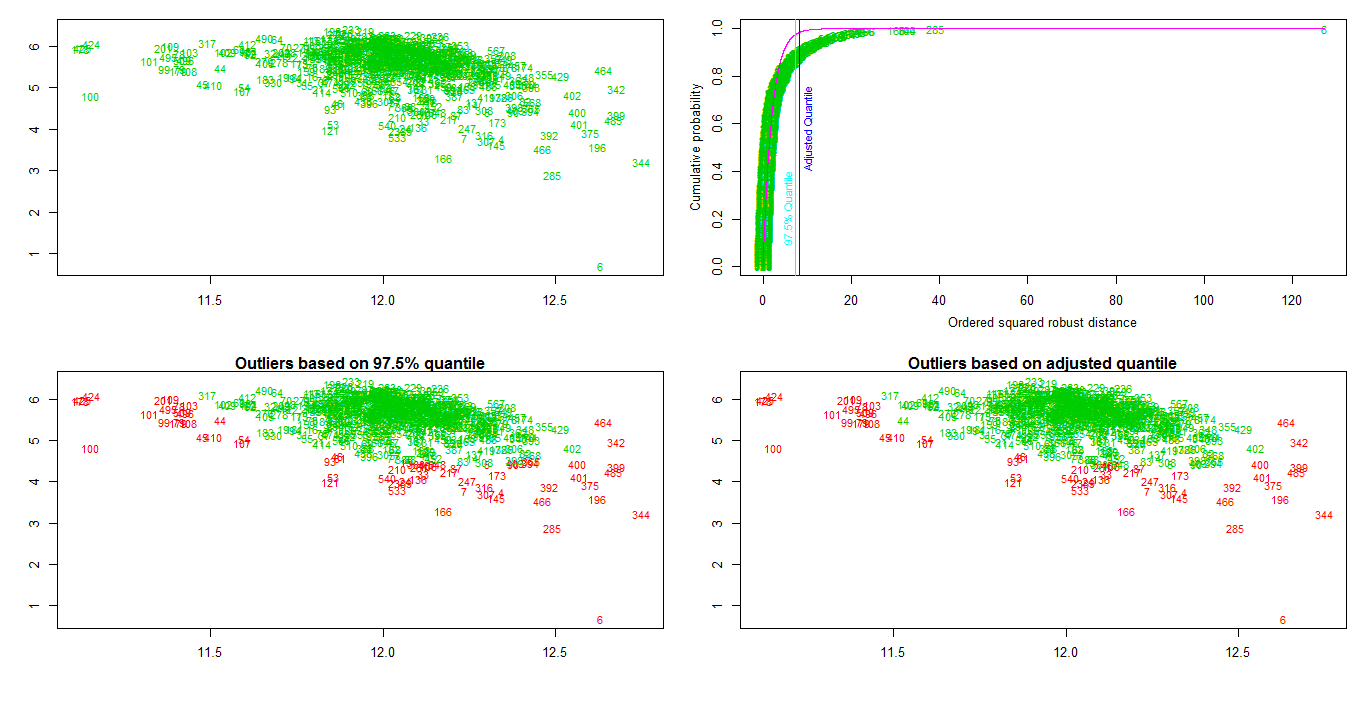


En las dos anterior graficas se puede concluir que en le caso de aplicar MAD en el Average Price elimina lecturas que en general se presentan y no son necesarias eliminarlas. En el caso de aplicar MAD al Seat Availability se eliminan valores un poco mas acorde a lo que llamariamos valores atipicos y genera una regresion mas adecuada para este periodo teniendo en cuenta el la anticipacion de compra es de un dia y los clientes estan dispuestos a pagar mas dinero para poder realizar su viaje

Adjusted Squared Plot

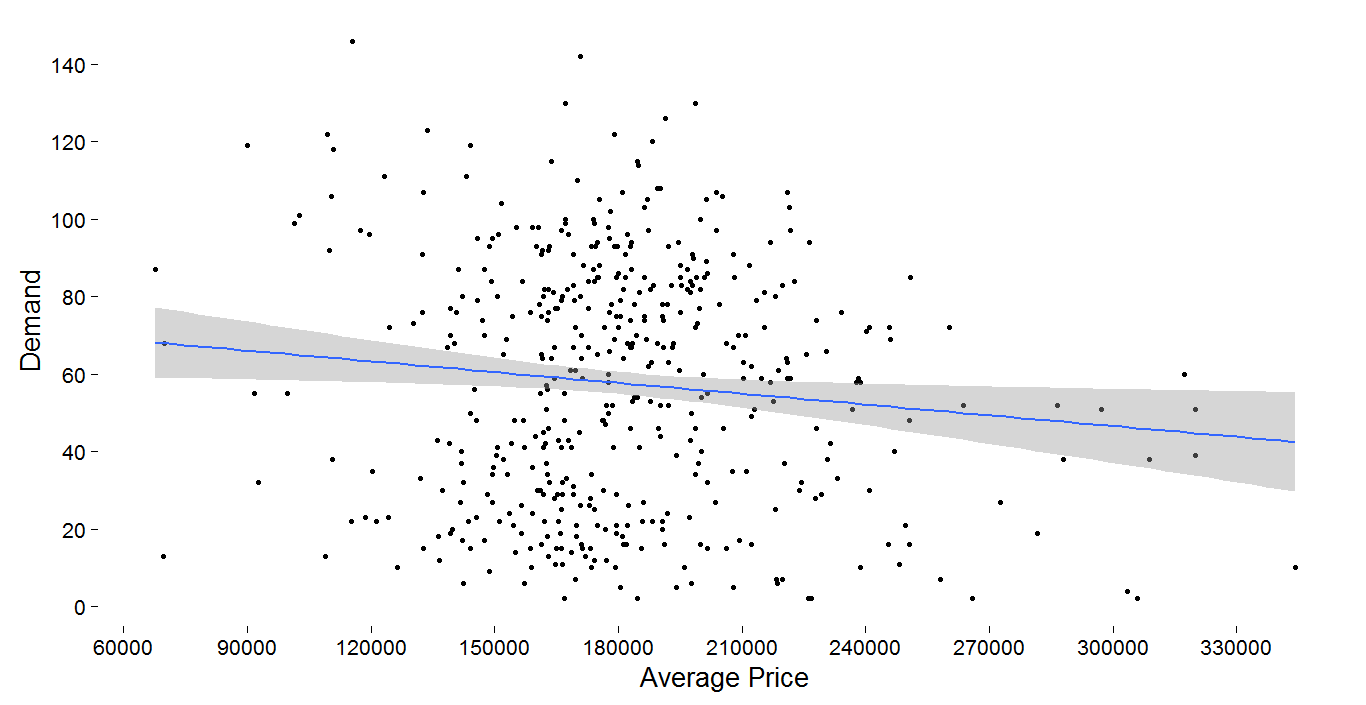
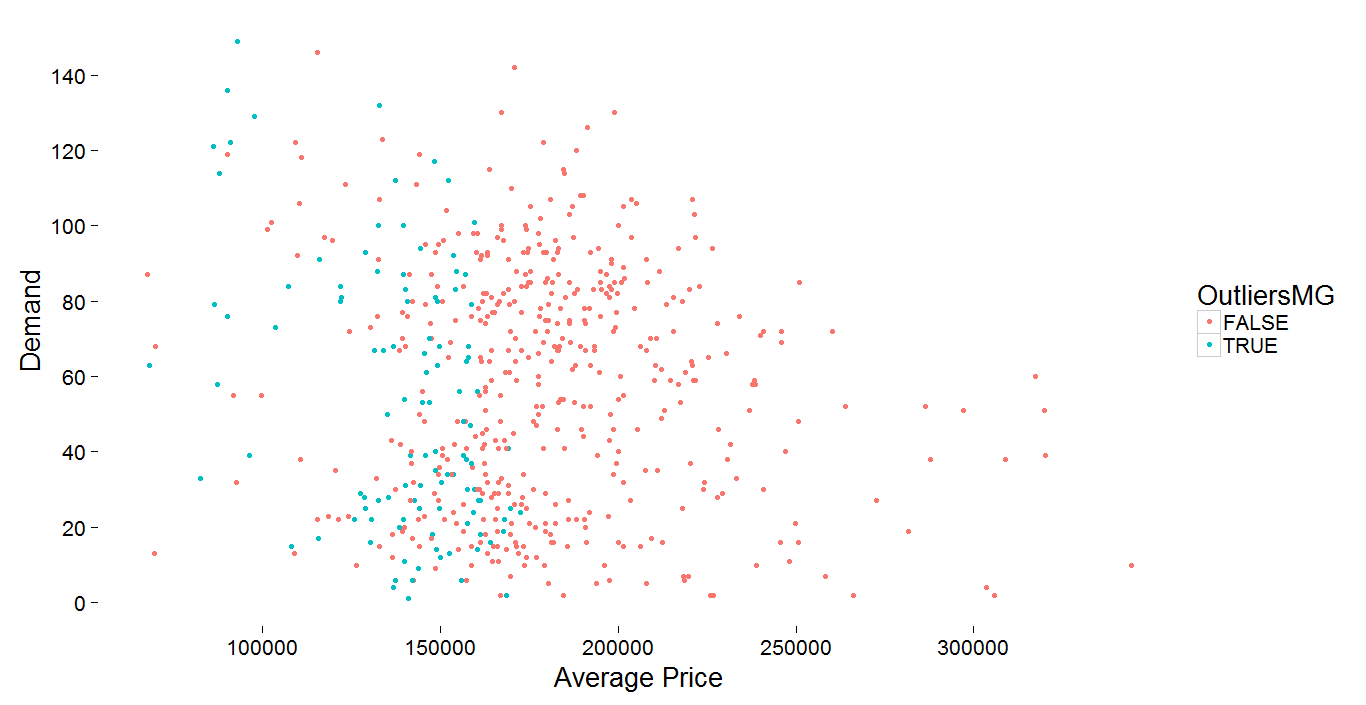
Este metodo multivariado utiliza las dos variables a la vez. Para poder cumplir el supuesto de normalidad, se aplico logaritmo a ambas variables.





Este metodo fue mas efectivo y permitio corregir la regresion, que como se menciono antes para este dia de anticipacion de compra los clientes estan dispuestos a pagar mas dinero para poder viajar.

Multivariate Guassian Distribution



En este metodo se eliminaron muchos puntos, lo que distorsiona el analisis

Se realizaron varias pruebas para diferentes OD Locales y diferentes anticipaciones de compra, se concluyo en primer lugar que las variables no cumplian el supuesto de normalidad pero que es necesario poder eliminar los valores atipicos para poder tener estimaciones mas adecuadas por lo que se definio que el modelo mas adecuado para eiliminar valores atipicos es el modelo de Adjusted Squared Plot.

### Construct Data

### Integrate Data

### Format Data

## Modelling

### Select Modelling Techniques

### Generate Test Design

### Build Model

### Asses Model

## Evaluation Phase

### Evaluate Results

### Process Review

### Determine Next Stops

## Deployment

### Deployment Plan

### Plan Monitoring and Maintenance

### Translation for Publication