

Introducción al Análisis Exploratorio y Preparación de los Datos

Maestría en Ciencia de Datos

Profesores:

Norha M. Villegas, Ph.D. (nvillega@icesi.edu.co)

Christian Urcuqui (ccurcuqui@icesi.edu.co)

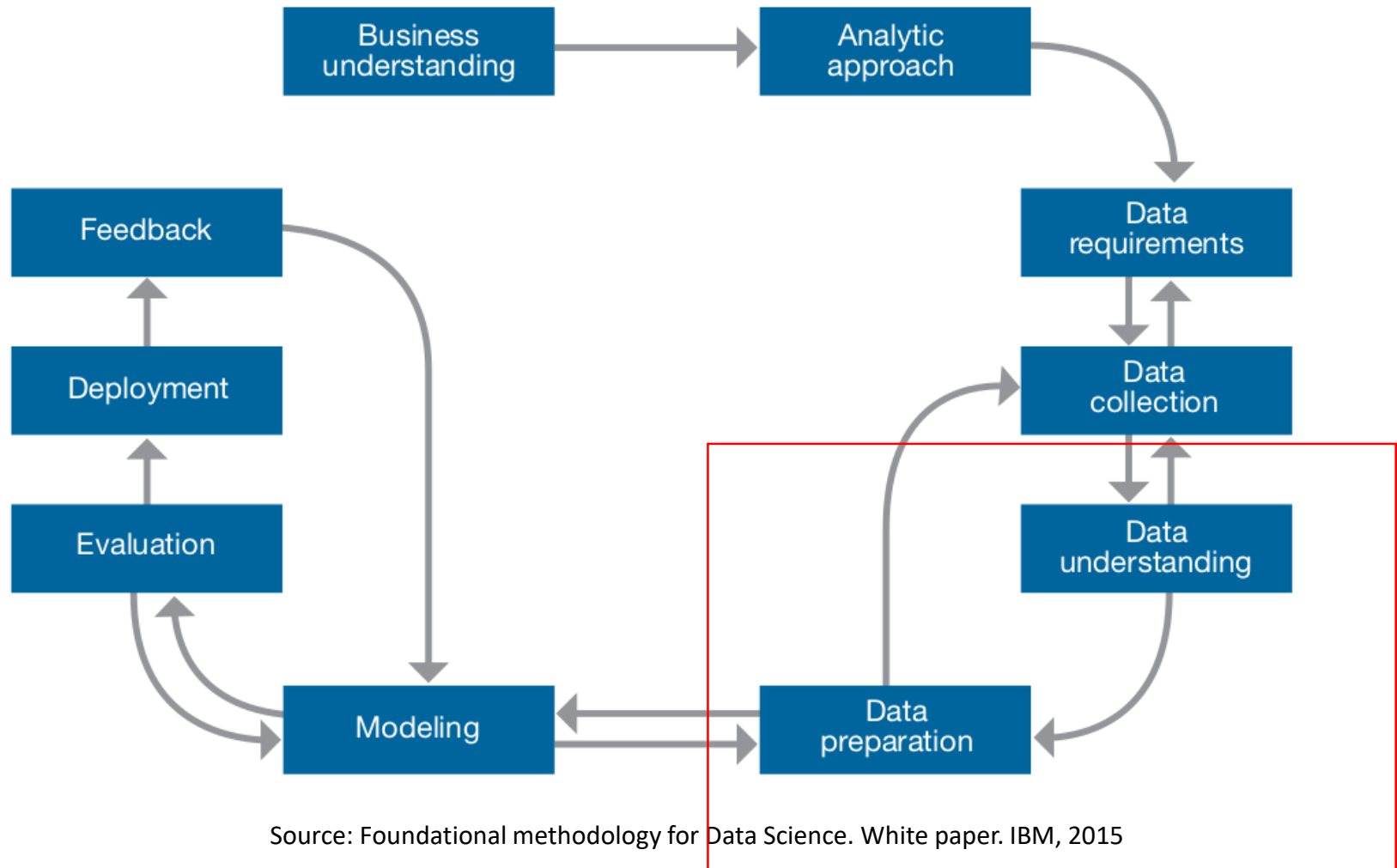
Universidad Icesi

Agenda para hoy

- Introducción a la analítica
 - Tipos de análisis
- Análisis exploratorio (y descriptivo)
 - Variables numéricas
 - Variables categóricas
- Limpieza y preparación de datos
- Aplicación de conceptos a un caso de negocio

Introducción a la analítica de datos

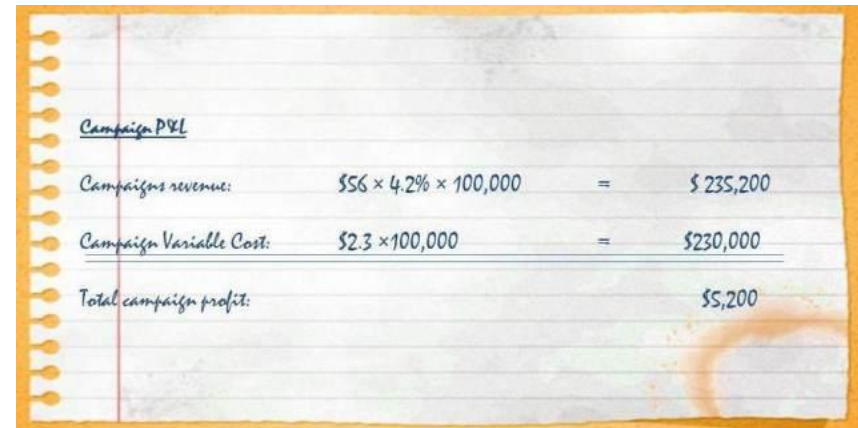
Ciclo de vida de la analítica



Entendiendo el ciclo de vida de la analítica: Marketing Analytics

- Usted es el director o directora de analítica de una tienda de comercio electrónico llamada DresSmart Inc. que se especializa en ropa.
- El director de mercadeo le informa que los miembros de la junta han aumentado los presupuestos de venta, al tiempo que han disminuido el presupuesto para invertir en campañas de mercadeo.
- Estrategia última campaña
 - Se enviaron correos físicos con el catálogo de productos a 100.000 clientes de la base de datos que tiene aproximadamente 2 millones de clientes.
 - La tasa de respuesta fue del 4.2%
 - Se hizo seguimiento a través de email y SMS a los clientes que recibieron el catálogo físicamente

Fuente: <http://ucanalytics.com/blogs/marketing-analytics-retail-case-study-part-1/>



<u>Campaign P&L</u>			
Campaign revenue:	$\$56 \times 4.2\% \times 100,000$	=	\$ 235,200
Campaign Variable Cost:	$\$2.3 \times 100,000$	=	\$230,000
Total campaign profit:			\$5,200

¿Cuál es el reto para la próxima campaña?

¿Cómo podemos ayudar desde la analítica de datos?

Tipos de preguntas en analítica (tipos de análisis)

- Análisis exploratorio
- Análisis inferencial
- Análisis predictivo
- Análisis causal
- Análisis mecanico

Menor dificultad



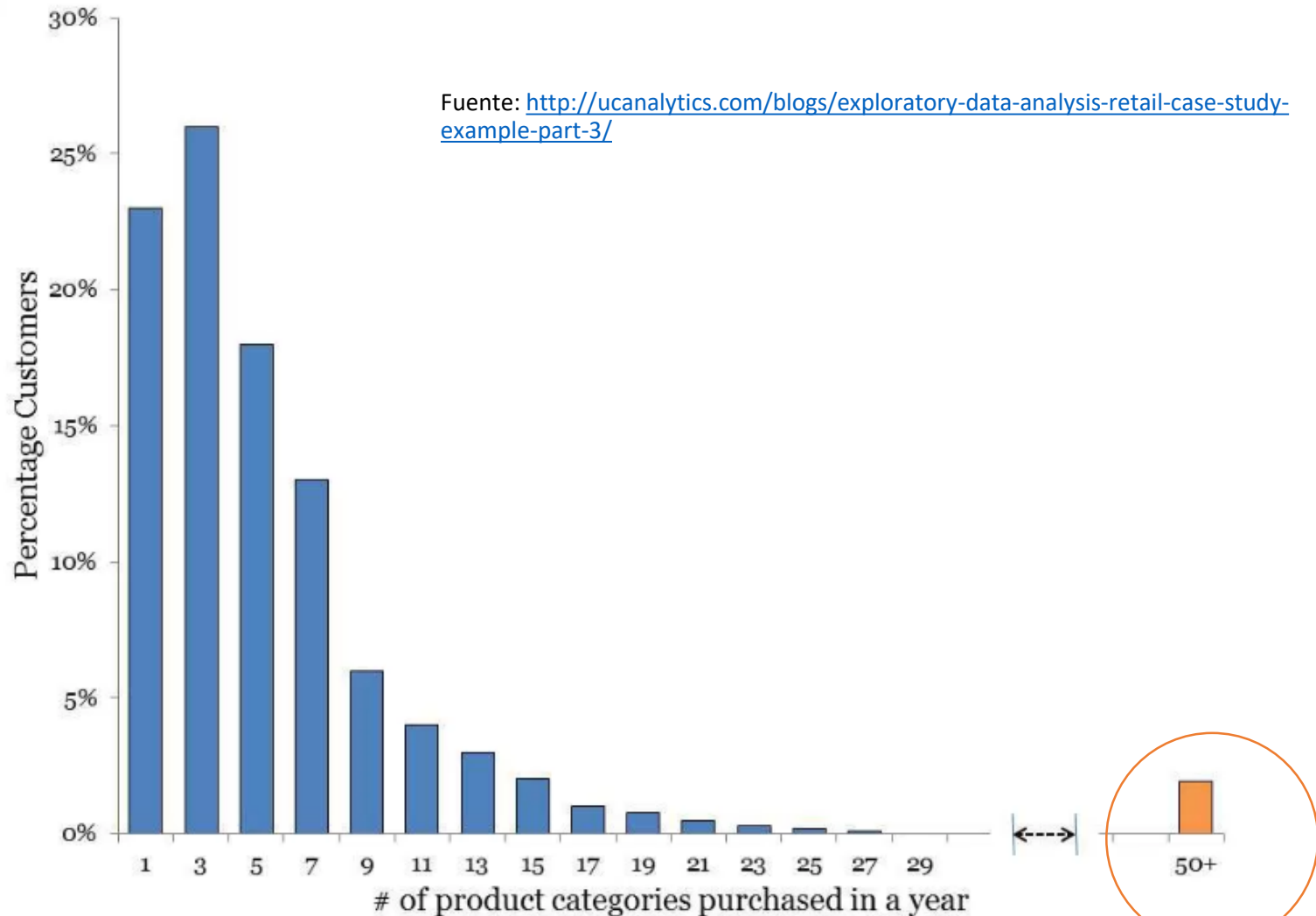
Mayor dificultad

Análisis exploratorio (EDA)

¡Entendiendo los datos!

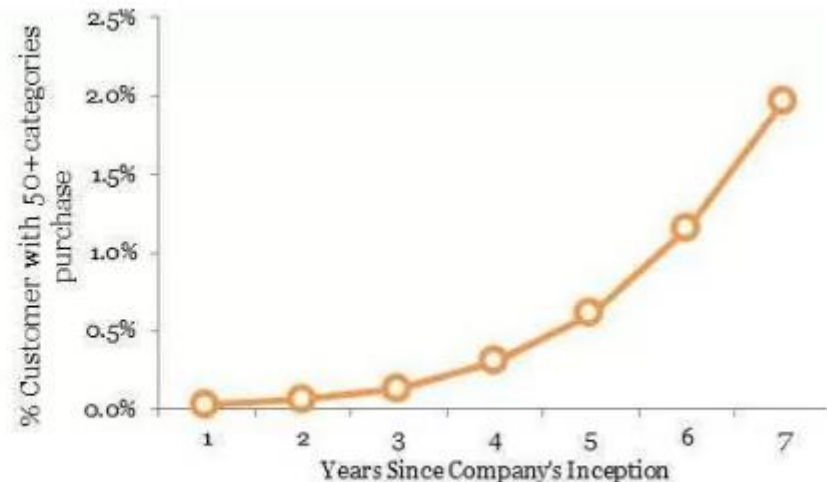
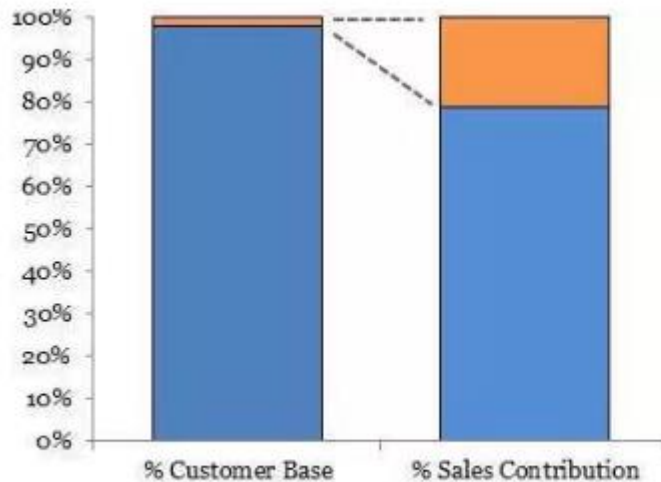
- Objetivos:
 - Entender la estructura del conjunto de datos
 - Detectar errores y deficiencias en los datos
 - Encontrar relaciones entre las variables
 - Identificar modelos candidatos para el análisis de los datos
- Primer tipo de análisis que se realiza
- Diferentes abstracciones de los datos
- Incluye el análisis descriptivo (e.j., BI tradicional), pero no modelos estadísticos formales, inferencias o predicciones
- Lo que se observa no puede generalizarse

Análisis exploratorio para DresSmart Inc.



Exploratory data analysis – marketing analytics case study (retail)

Análisis exploratorio



Exploratory data analysis

¿Quiénes son estos clientes?

¿Qué tipo de productos compran estos clientes que parecen tener un comportamiento interesante para la compañía?

R/ Usualmente los mismos estilos en muchas diferentes tallas

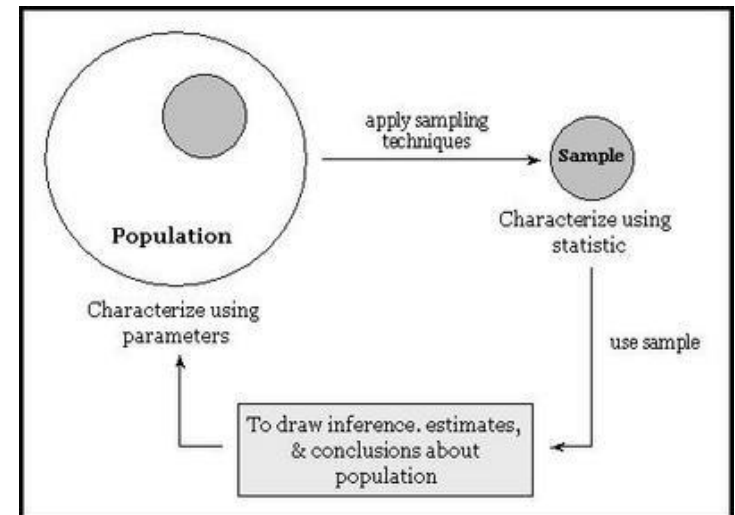
¿Podríamos concluir algo acerca de estos clientes? ¿qué tipo de cliente son?

¿Cómo impactan estos hallazgos la estrategia para la campaña de mercadeo?

Análisis inferencial

¡Concluamos acerca de la población!

- Objetivos:
 - Analizar una muestra relativamente pequeña para decir algo acerca de la población a la que pertenece la muestra
- Se basa en modelos estadísticos
- Verificar una hipótesis y concluir sobre la población con base en el análisis de la muestra
- Determinar la probabilidad de que se presenten características en la población con base en las características de la muestra
- La selección de la muestra es crucial



Fuente:

<http://www.discover6sigma.org/post/2005/12/statistics-simplified/>

Análisis predictivo

- Objetivo
 - Predecir hechos futuros o desconocidos a partir del análisis de datos actuales e históricos
- Aplican técnicas de la estadística, minería de datos, aprendizaje automático (machine learning), inteligencia artificial

Análisis predictivo

Business Analytics: From Descriptive To Predictive...

...then use Predictive Analytics to build predictive models and actionable recommendations at the individual consumer and store levels



Mia

Predictive Answers

- How many times is Mia likely to shop our store over Christmas?
- What promotions is Mia likely to use?
- What's the likelihood that Mia will buy Product Y?
- What's probability Mia will buy product Z when she buys Product Y?
- What's the profit potential of Mia over the next 2 years?

Recommendations

- What's best price to get Mia to buy private label cookies?
- What are the best promotions, and on what products, to get Mia to visit our stores 2 additional times a month?
- What new product introductions should we recommend to Mia?
- What private label products have the best chance of Mia buying?

Hindsight = Demographics,
Preferences, Behaviors



Foresight = Personalized,
Predictable, Actionable, Measurable

EMC

Fuente: DELL EMC - https://infocus.emc.com/william_schmarzo/business-analytics-moving-from-descriptive-to-predictive-analytics/

Análisis predictivo

Business Intelligence answers...	Predictive Analytics answers...
When did customer X last visit the store?	What's the probability that customer X will visit the store tomorrow?
How many customers visited the store over the past week?	How many customers are likely to visit the store next week?
How much revenue did store X generate last Christmas?	How much revenue will be generated by store X next Christmas?
What's the revenue trend for a particular product line over the past 12 months?	What's the projected revenue trend for a particular product line over the next 12 months?
How many leads did our last marketing campaign generate?	How many leads will our next marketing campaign generate?
From what cities, geographies, and zip codes do my highest revenue customers come?	What are the revenue potential of my customers by City, Geography and Zip Code?
What are the most profitable products for my Diamond and Platinum customers over the past 12 months?	What could be the most profitable products for my Diamond and Platinum customers over the next 12 months?
	What behaviors are indicative of a new customer's likelihood to advance to the Platinum or Diamond levels?
	What customers are most likely to be interested in the new product given their product purchase behaviors and interests?
	What's the estimated impact on revenue, customer spend, and store traffic with the addition of a new product or product category?
	What is the store traffic and spend impact featuring a promotion with Reba McIntire versus Air Supply?
	What's the lifetime value of a customer considering their visits, spend and advocacy ratings?
	What are the correlations between product purchases and non-purchase activities that drive the most profitable customer visit?
	What are unusual customer usage and behavioral patterns that may indicate attrition, up-sell, cross-sell or fraud?

Otros tipos de análisis

- Análisis causal
 - Objetivo: Identificar qué sucede con una variable cuando se cambia otra
- Análisis mecánico
 - Objetivo: Entender qué cambios en qué variables dan lugar a cambios en otras variables
 - Más usado en el análisis de fenómenos físicos

El entendimiento de los datos es fundamental para la preparación de los mismos y para la selección de modelos de analítica de datos

Análisis exploratorio

El dataset (ejercicio práctico)

- Descargar del Moodle el dataset custdata.tsv
- Cargarlo en R en un DataFrame
- Verifique que se crea el DataFrame y que contiene 1.000 observaciones con 11 variables

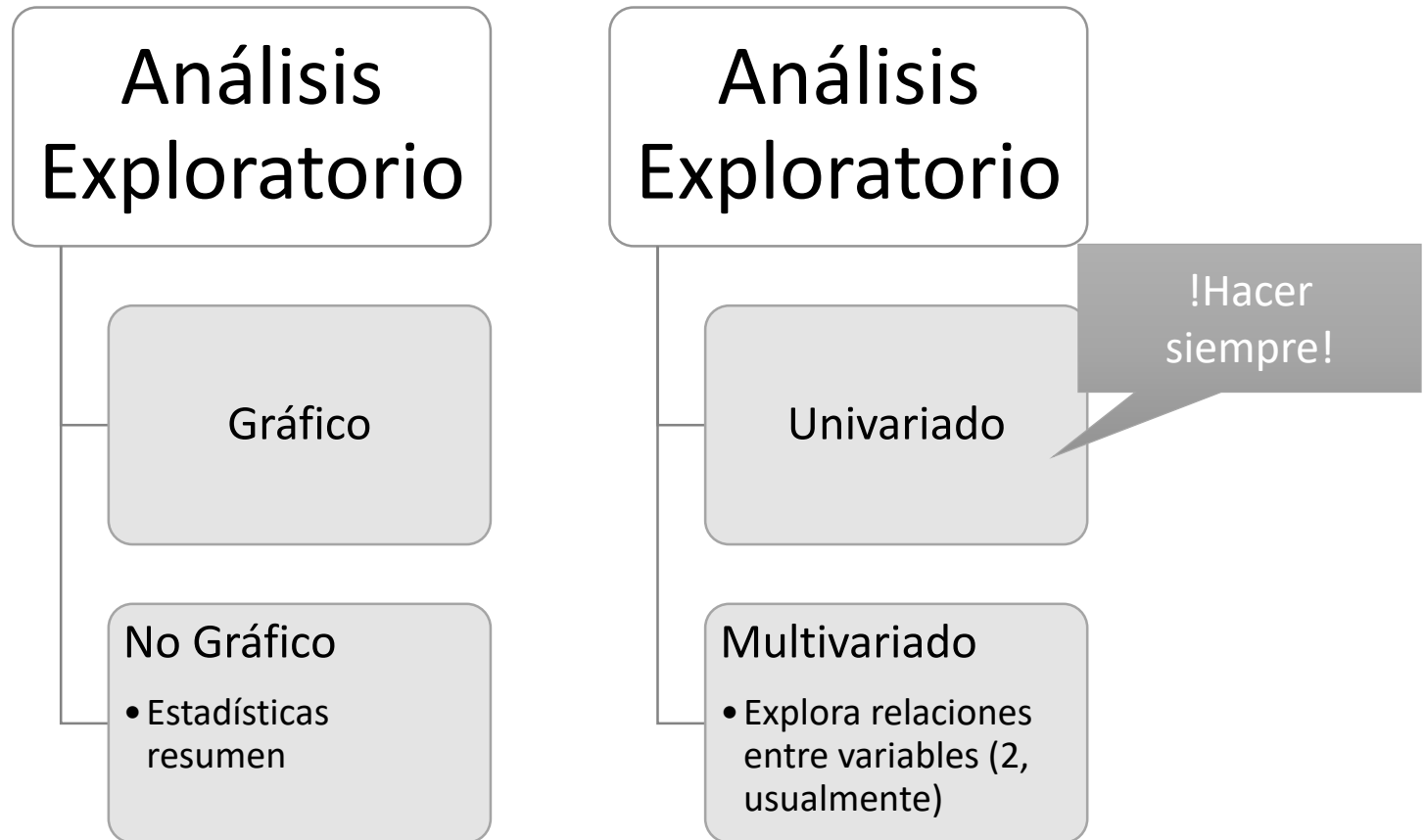
¿Cómo podríamos explotar este dataset para ayudarle al negocio a lograr su objetivo de ofrecer pólizas de seguro de salud complementarias?

¿Por dónde empezamos?

Estadísticas resumen y de conteo (función Summary – ejercicio práctico)

- Str: Visualización de la estructura de los datos
- head, tail: visualizando los primeros y últimos datos del dataset
- Summary: primera aproximación a los datos
- **Actividad**
 - Aplicar la función summary al dataframe
 - En parejas analizar la información que se obtiene para cada una de las variables
 - Escribir lo que se observa en los datos a partir de la información suministrada por la función summary

Análisis exploratorio (EDA)



Exploración variables numéricas: análisis no gráfico

Estadísticas resumen, medidas de tendencia central

Exploración de variables: estadísticas resumen

- Numéricas: funciones estadísticas de resumen
 - Valor mínimo
 - Cuartiles
 - Mediana
 - Media
 - Valor máximo
 - Útiles para
 - Entender la distribución de la muestra para aproximarse a la distribución de la población
 - Detectar datos atípicos (outliers, valores fuera de las distribuciones más comunes)
- Categóricas: visualizaciones
 - Útiles para entender la distribución de los datos y las relaciones entre variables

Exploración de variables numéricas: estadísticas resumen (cont.)

- Medidas de tendencia central (valores medios típicos)
 - Media (promedio)
 - Mediana
- Analicemos la media y la mediana para las variables numéricas de nuestro dataset. ¿Qué aprendemos acerca de los datos?

Exploración variables numéricas: análisis gráfico

Revisión gráfica las distribuciones estadísticas (rangos, subpoblaciones, valores inválidos, valores atípicos)

Análisis gráfico

- Identificar el valor pico de las distribuciones
- Identificar si la distribución es unimodal o multimodal
- Identificar qué tan normal o lognormal son los datos
- Identificar la variación de los datos, ¿se concentran en cierto intervalo o categoría (en el caso de las categóricas)?

Exploración de variables numéricas (cont.)

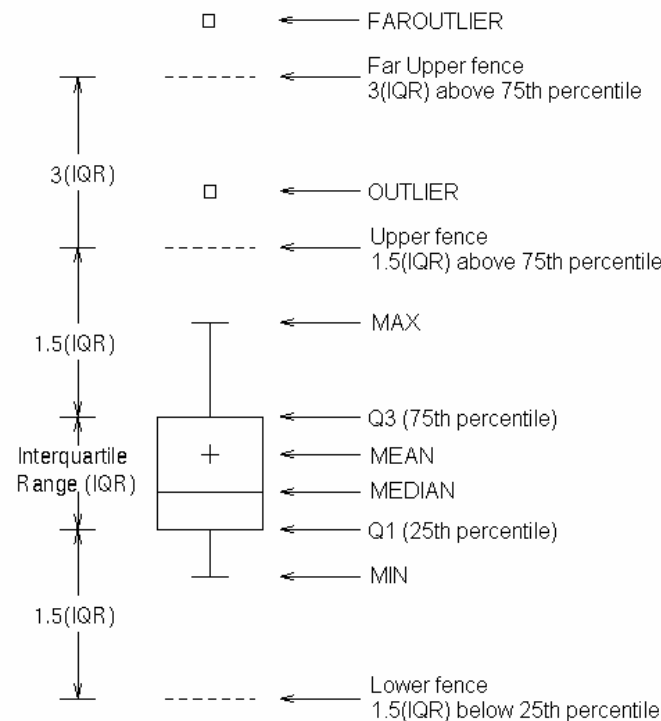
- Medidas de dispersión de los datos
 - ¿cómo son los rangos de las variables?
- Cuartiles y el resumen de 5 números
 - Los tres valores que dividen un conjunto de datos ordenado en cuatro partes
 - Nos ayudan a entender qué tan dispersos están los datos, qué tan diversos son
 - Entender si la mayoría de los datos se parecen o no a las medidas centrales (promedio, mediana)

```
> summary(custdata$age)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0	38.0	50.0	51.7	64.0	146.7

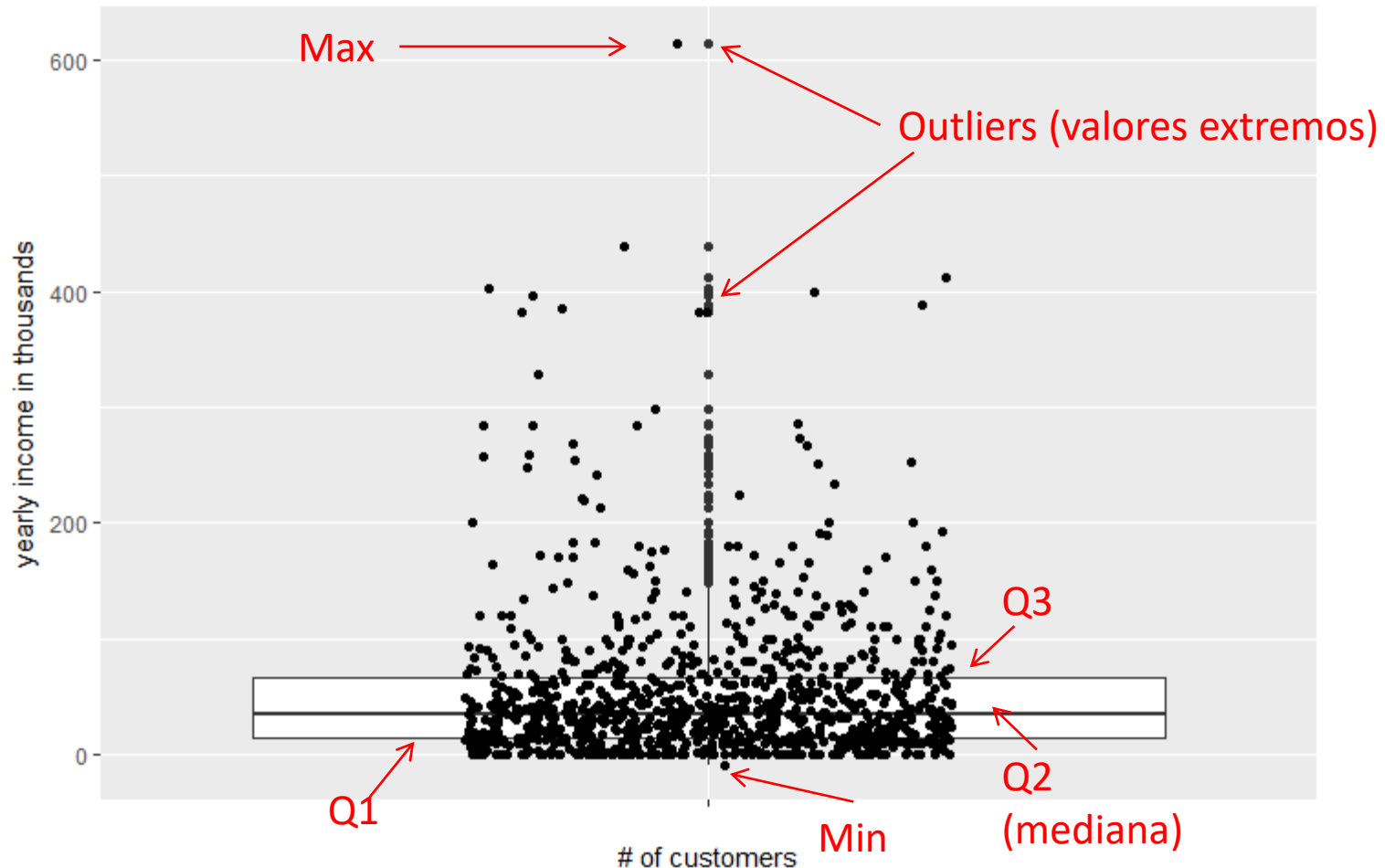


Visualización de variables numéricas: diagramas de cajas y bigotes



Visualización de variables numéricas: diagramas de cajas y bigotes

Visualizar niveles de dispersión de una variable



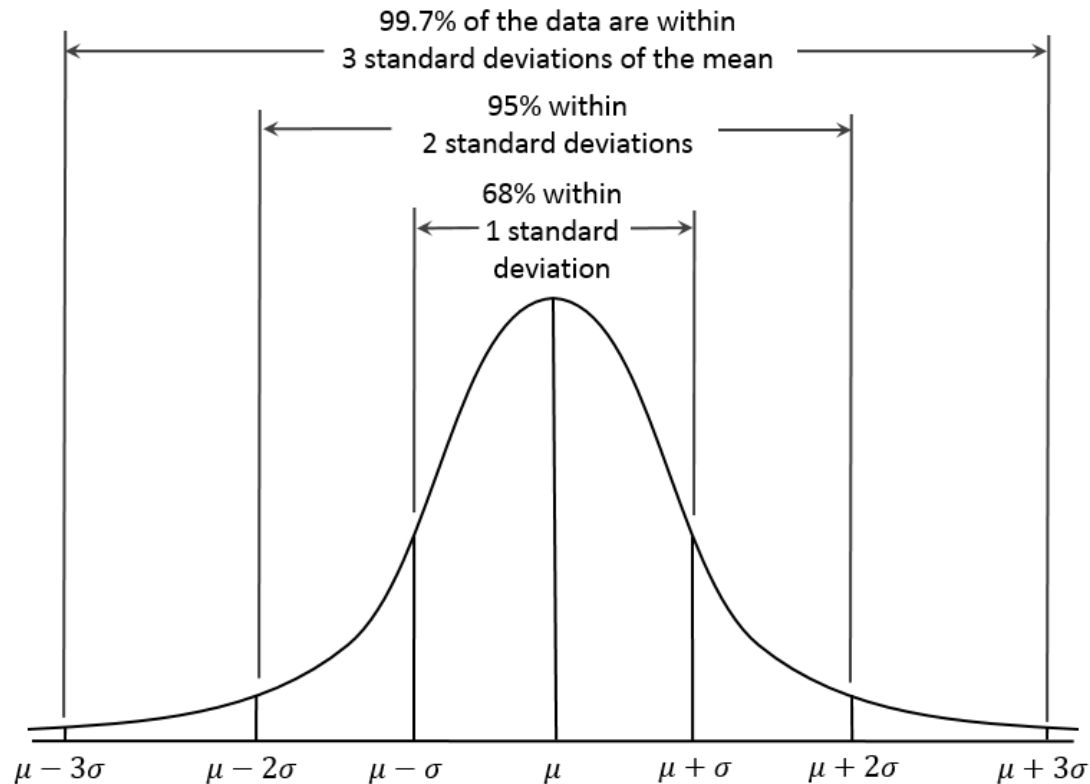
Visualización de variables numéricas: diagramas de cajas y bigotes

- Analicemos los cuartiles de las variables numéricas de nuestros datasets. ¿Qué encontramos?

```
custdata$income2<-custdata$income/100
```

```
ggplot(custdata, aes(x=" ", y=custdata$income2)) +  
  geom_boxplot() + geom_jitter(width = 0.2) + ylab("yearly income in hundreds") + xlab("# of customers")
```

Midiendo dispersiones: desviación estándar



Source: Wikipedia

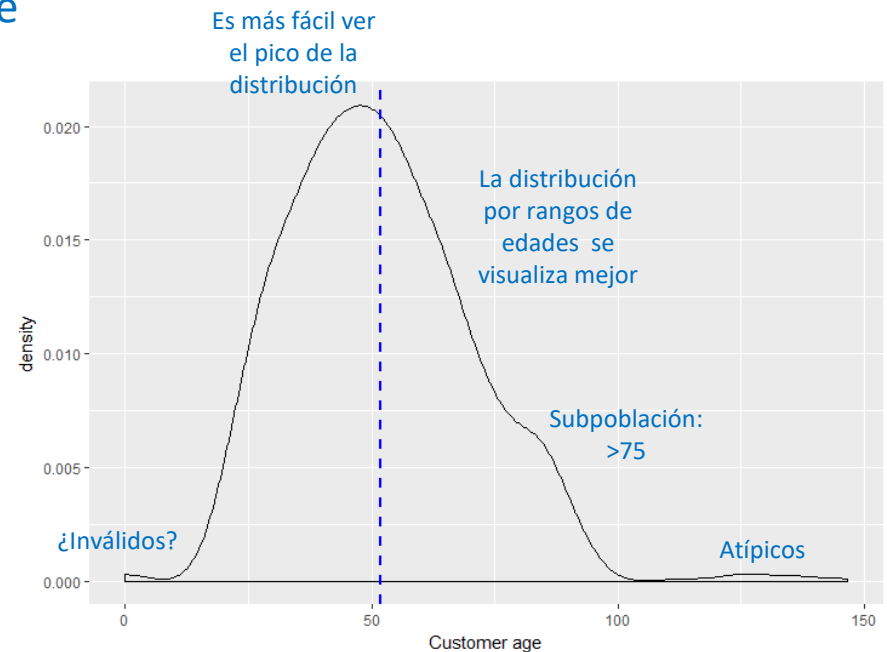
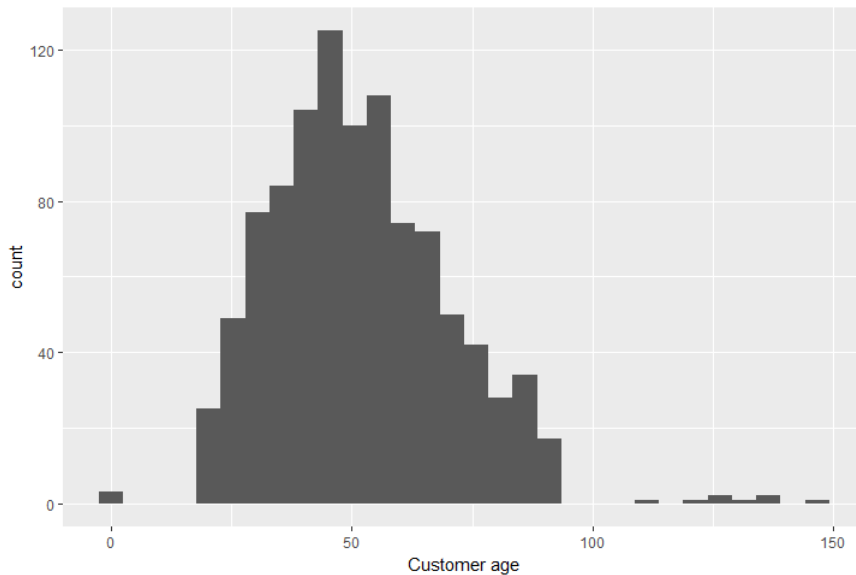
By Dan Kernler - Own work,

CC BY-SA 4.0, <https://commons.wikimedia.org/w/index.php?curid=36506025>

Visualización de variables numéricas: histogramas y diagramas de densidad

Visualizar niveles de dispersión de una variable, dónde se concentran las variables

age



```
> summary(custdata$age)
```

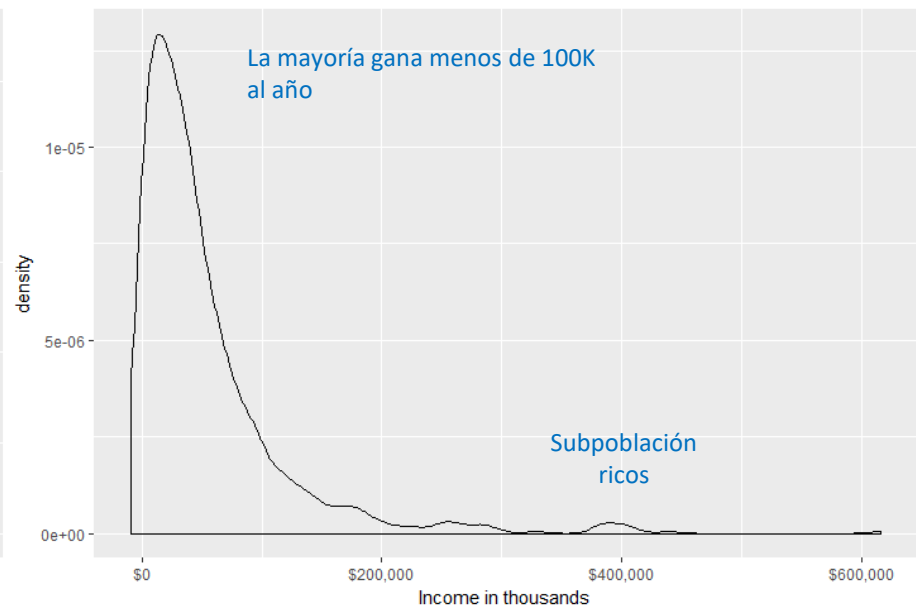
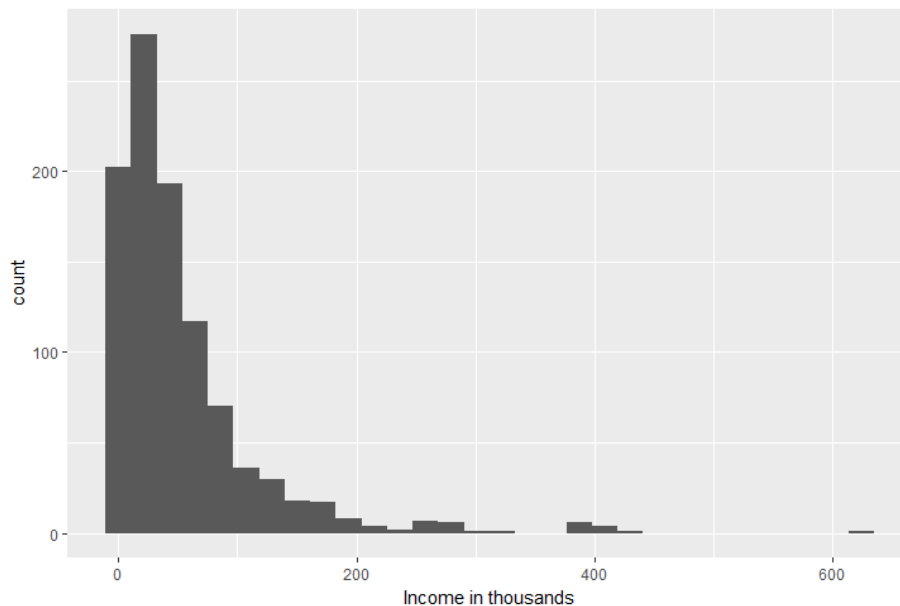
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0	38.0	50.0	51.7	64.0	146.7

¿Qué podemos decir de esta variable?

Visualización de variables numéricas: histogramas y diagramas de densidad

Visualizar niveles de dispersión de una variable

income



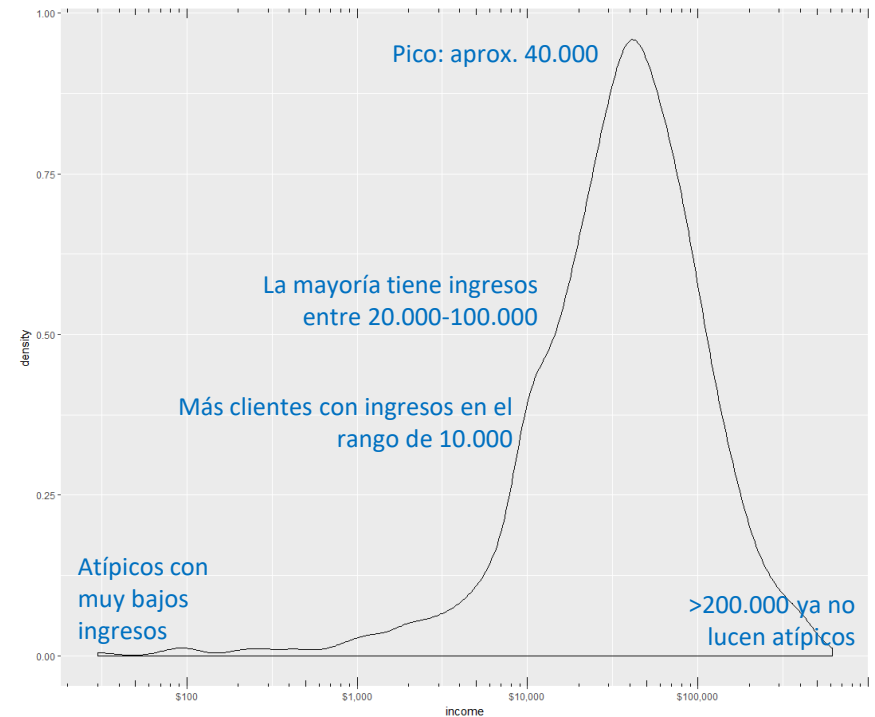
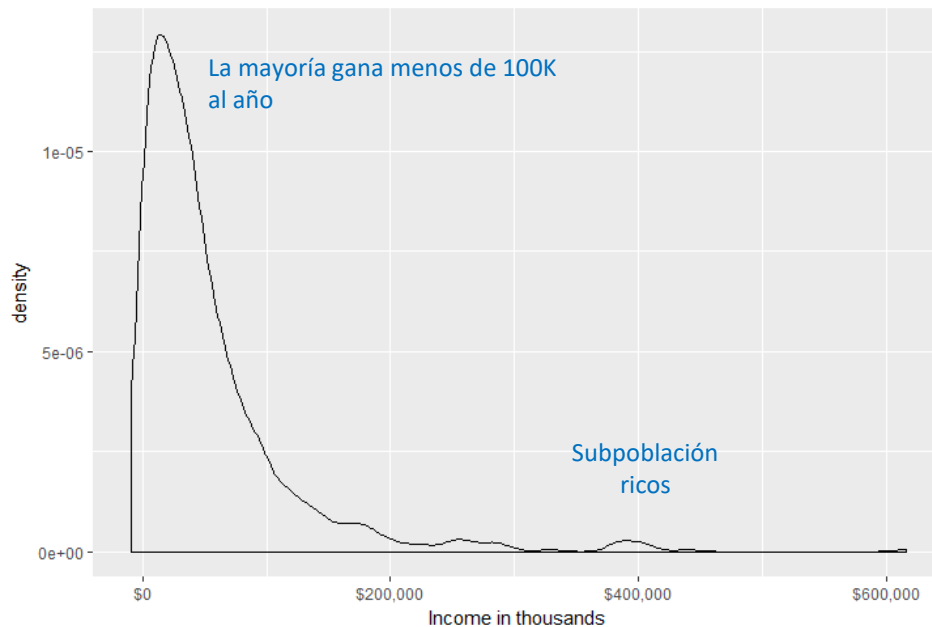
```
> summary(custdata$income)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-8700	14600	35000	53505	67000	615000

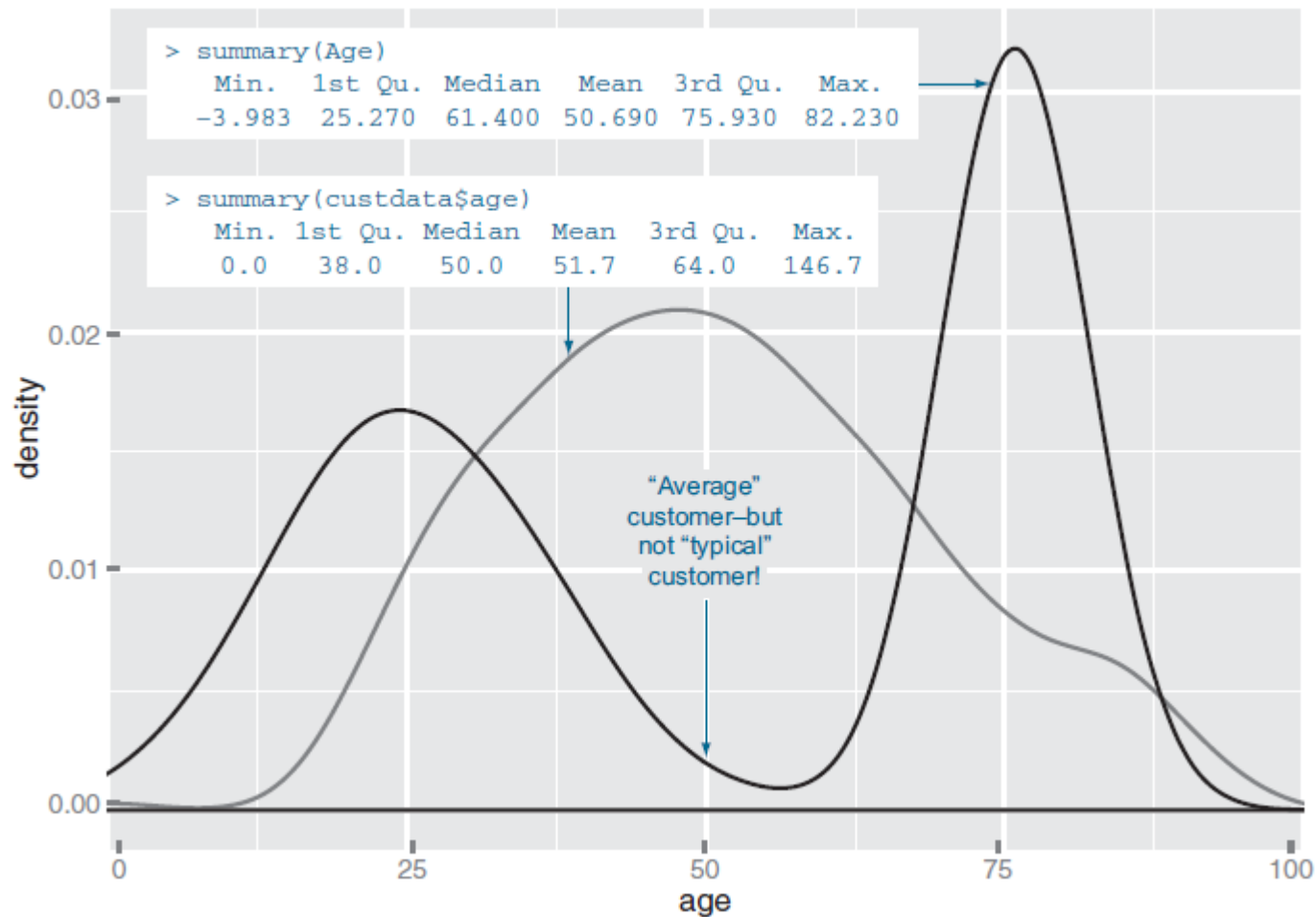
¿Qué podemos decir de esta variable?

Transformación de variables: preparación

income



Distribución unimodal vs. bimodal



Exploración variables categóricas

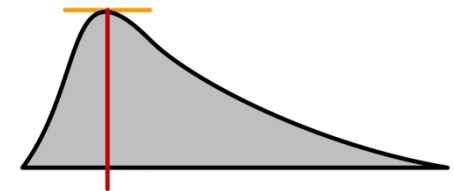
Exploración no gráfica de variables categóricas

- ¿Cuáles son variables categóricas en nuestro dataset de clientes?
- Este tipo de variables se examinan generalmente usando tablas de frecuencia para las categorías en lugar de estadísticas resumen.
- ¿Cómo podemos generar una tabla que nos muestre el número de clientes por género?

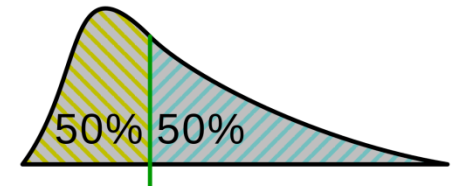
```
table(custdata$sex)  
prop.table(table(custdata$sex))
```

Exploración de variables categóricas

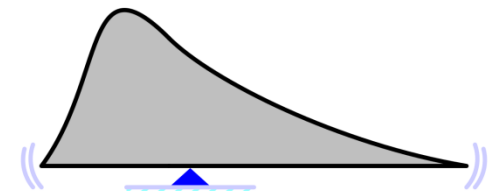
- La moda
 - Medida central usualmente aplicada a variables categóricas
 - Nos permite entender valores importantes de las variables categóricas. ¿Existen categorías que sean más importantes que otras?
 - Las variables pueden ser unimodales, bimodales o multimodales
 - ¿Cuál es la moda de las variables categóricas de nuestro dataset de customers?



mode



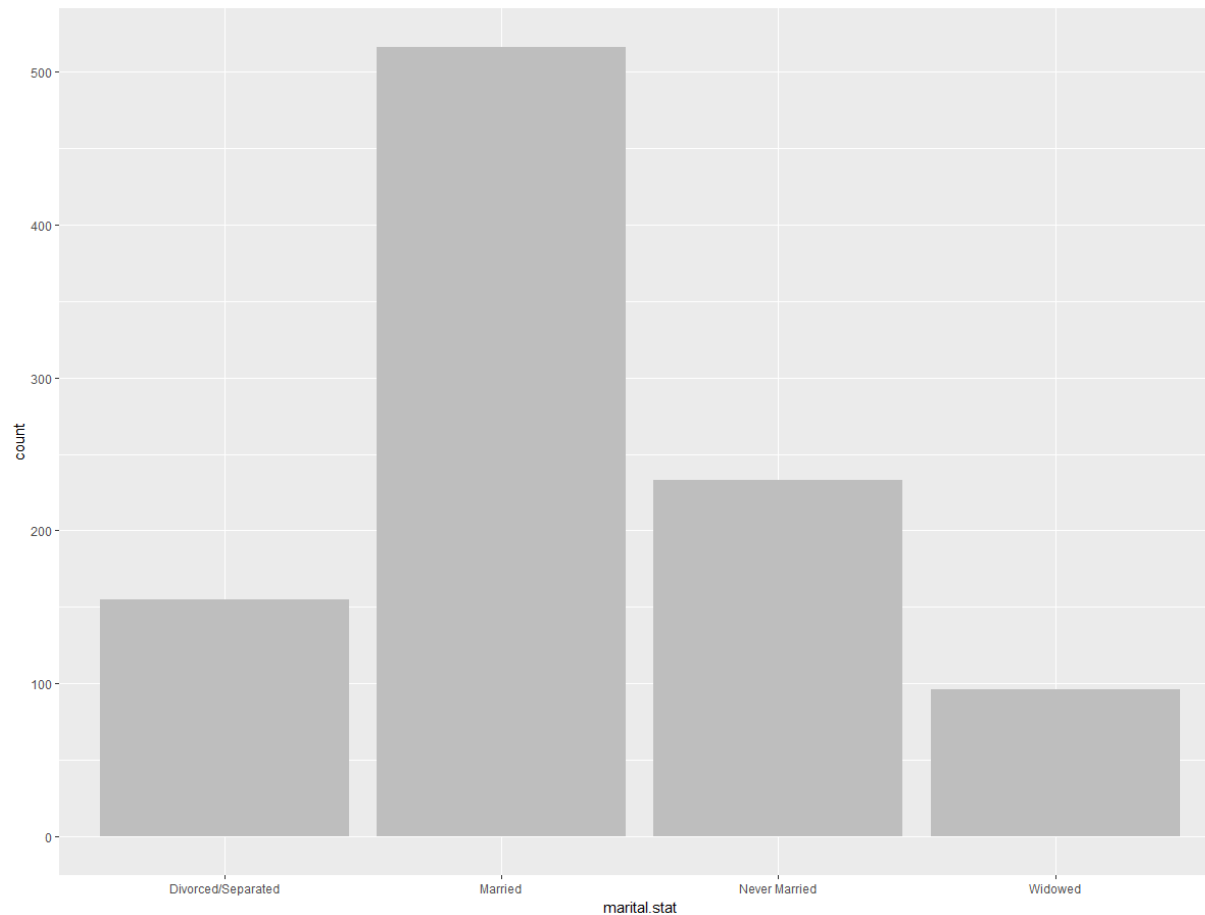
median



mean

Gráficos de barras: exploración gráfica de variables categóricas

- Compara las frecuencias absolutas o relativas de variables categóricas



Gráficos de barras: exploración gráfica de variables categóricas

- Compara las frecuencias absolutas o relativas de variables categóricas

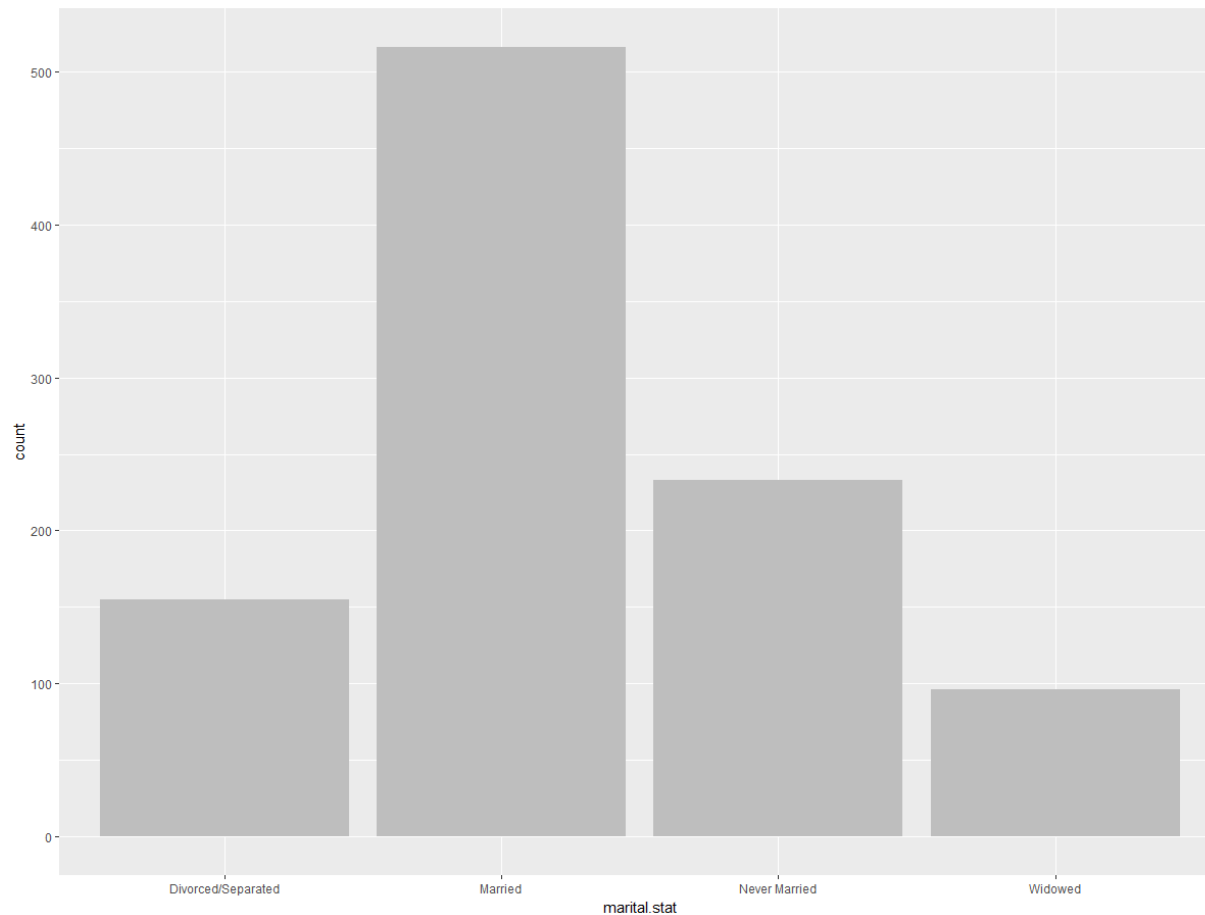


Gráfico de barras horizontal

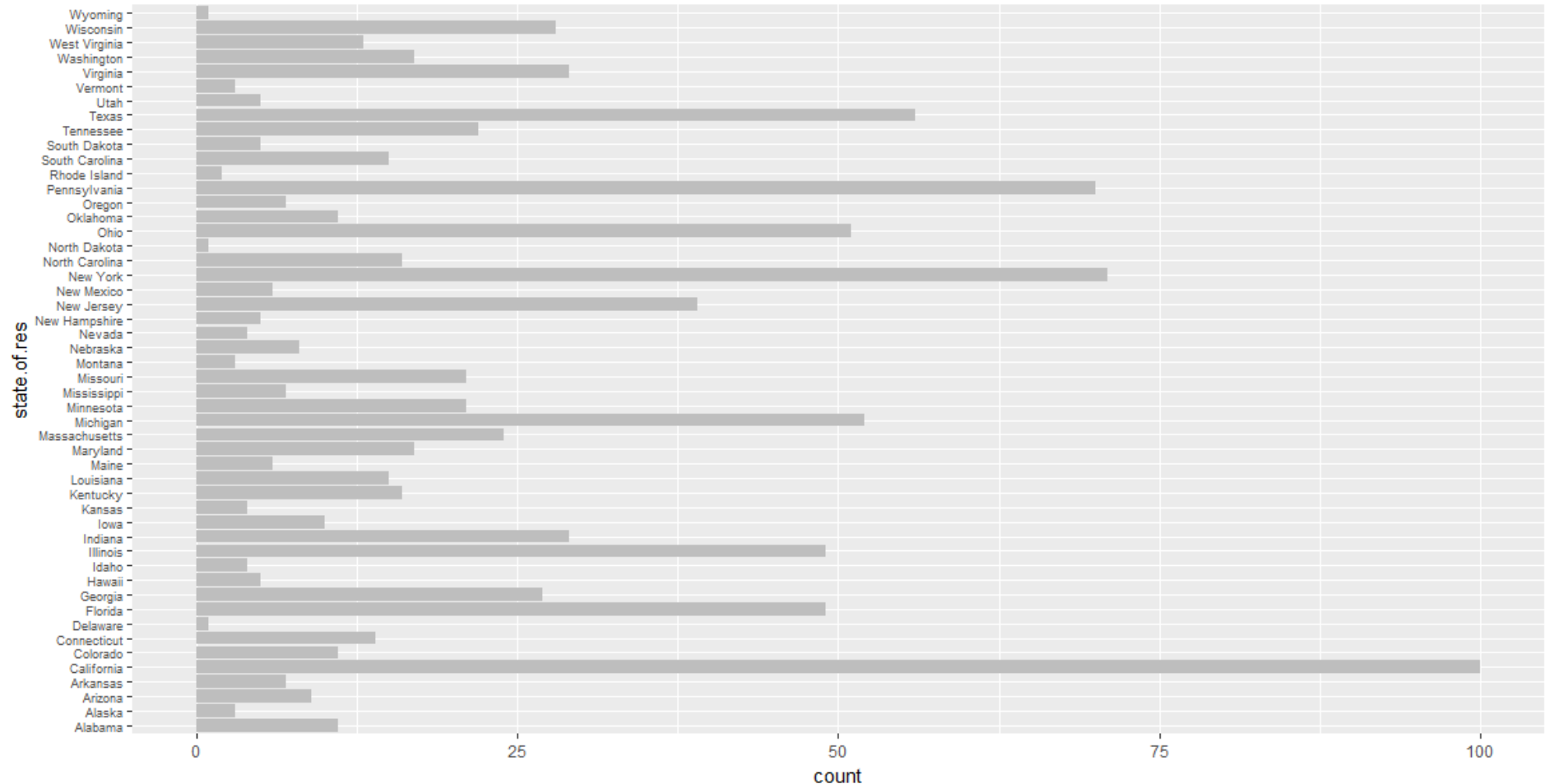
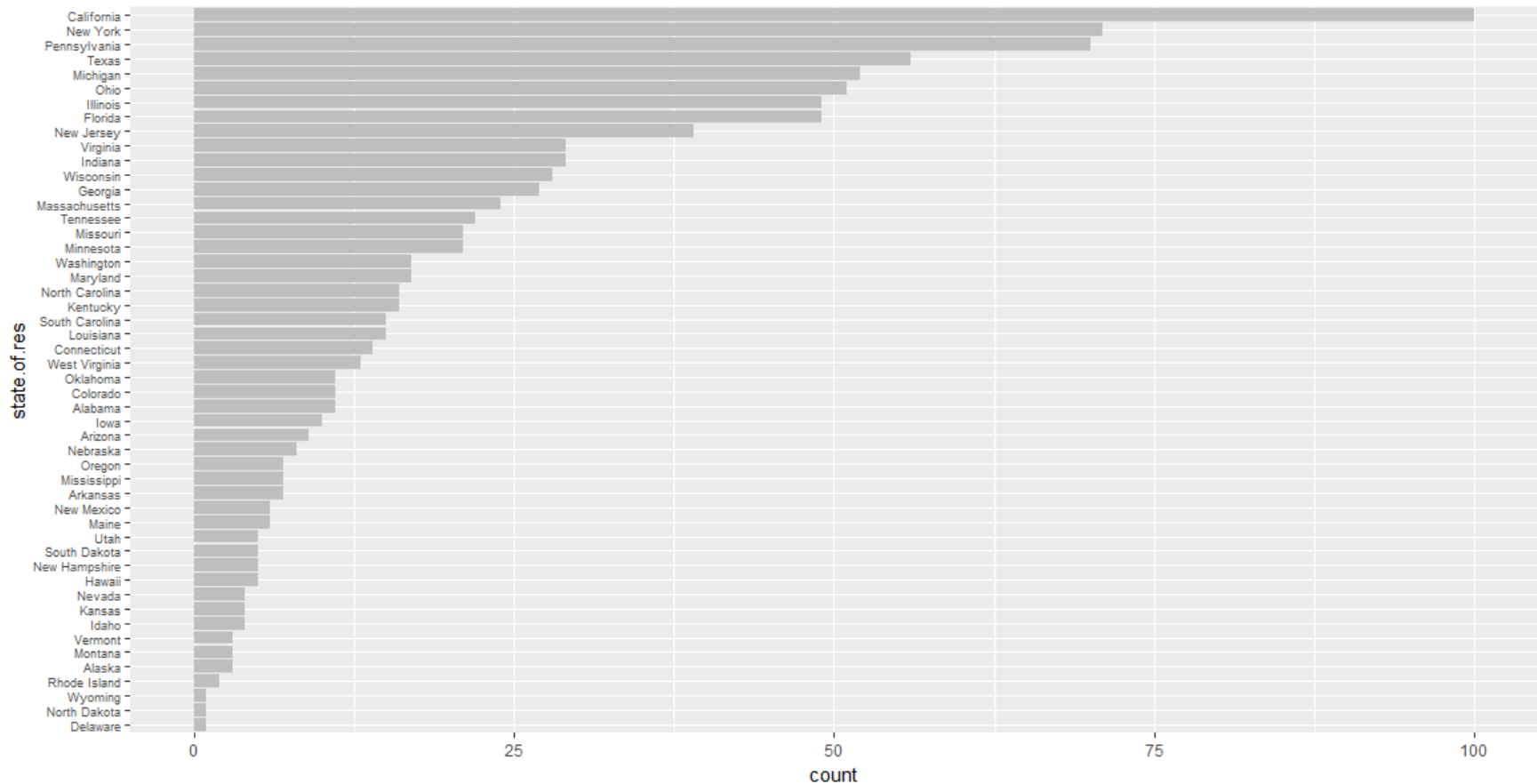


Gráfico de barras horizontal ordenado por frecuencias



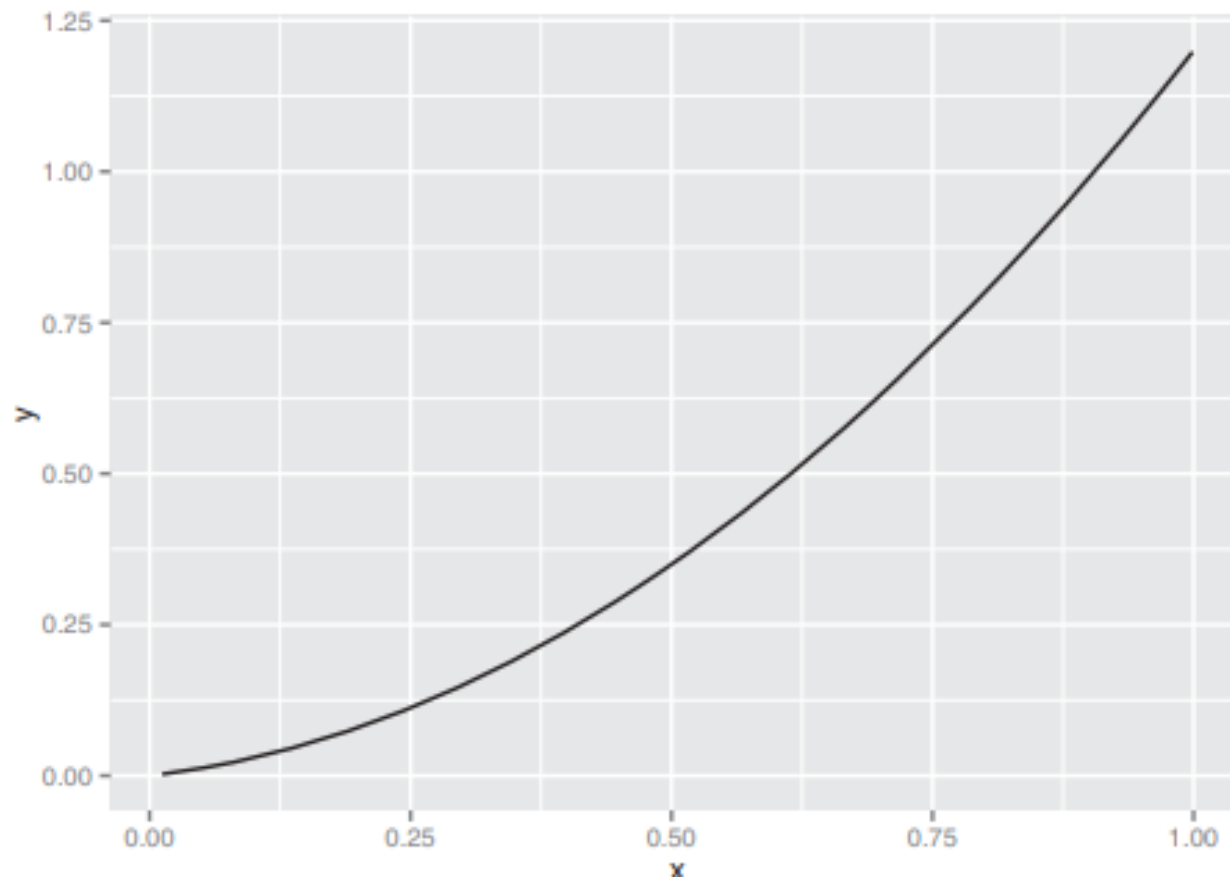
Análisis multivariado

Análisis multivariado: tipos de preguntas

- Nos permite explorar relaciones entre dos o más variables para ir determinando variables explicativas que sean candidatas para ser incluidas en el modelo
- Contestemos las siguientes preguntas
 - ¿Existe correlación entre edad e ingreso? ¿Qué tipo de relación? ¿Qué tan fuerte es?
 - ¿Existe alguna relación entre estado civil y el hecho de que la persona tenga o no un seguro de salud?
 - ¿Cómo lo haríamos?

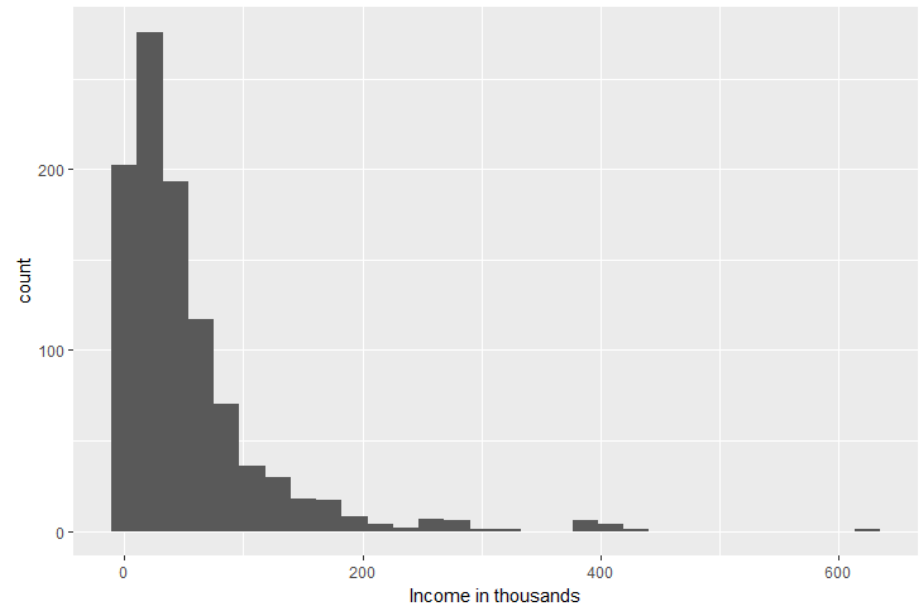
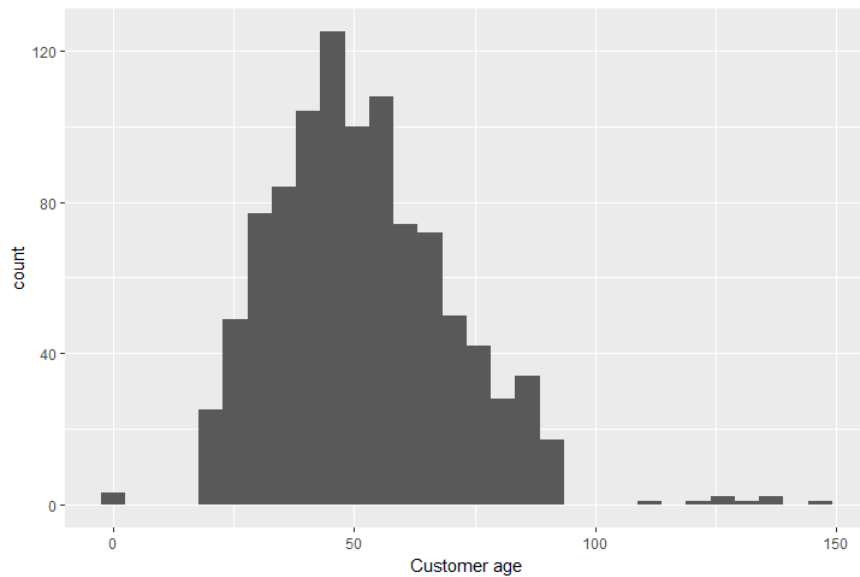
Gráfico de línea (geom_line)

- Relaciones limpias: a cada valor de x le corresponde un valor en y



Relaciones entre variables

- ¿Existirá una relación entre edad e ingreso?



Relaciones entre variables

- ¿Existirá una relación entre edad e ingreso?

```
custdata2 <- subset(custdata,  
  (custdata$age > 0 & custdata$age < 100  
    & custdata$income > 0))
```

Limpiamos un
poco el dataset

```
> cor(custdata2$age, custdata2$income) [1] -0.02240845
```

Calculamos la
correlación

¿Cómo interpretamos esta correlación?

Gráfico de dispersión (geom_point)

- ¿Existirá una relación entre edad e ingreso?

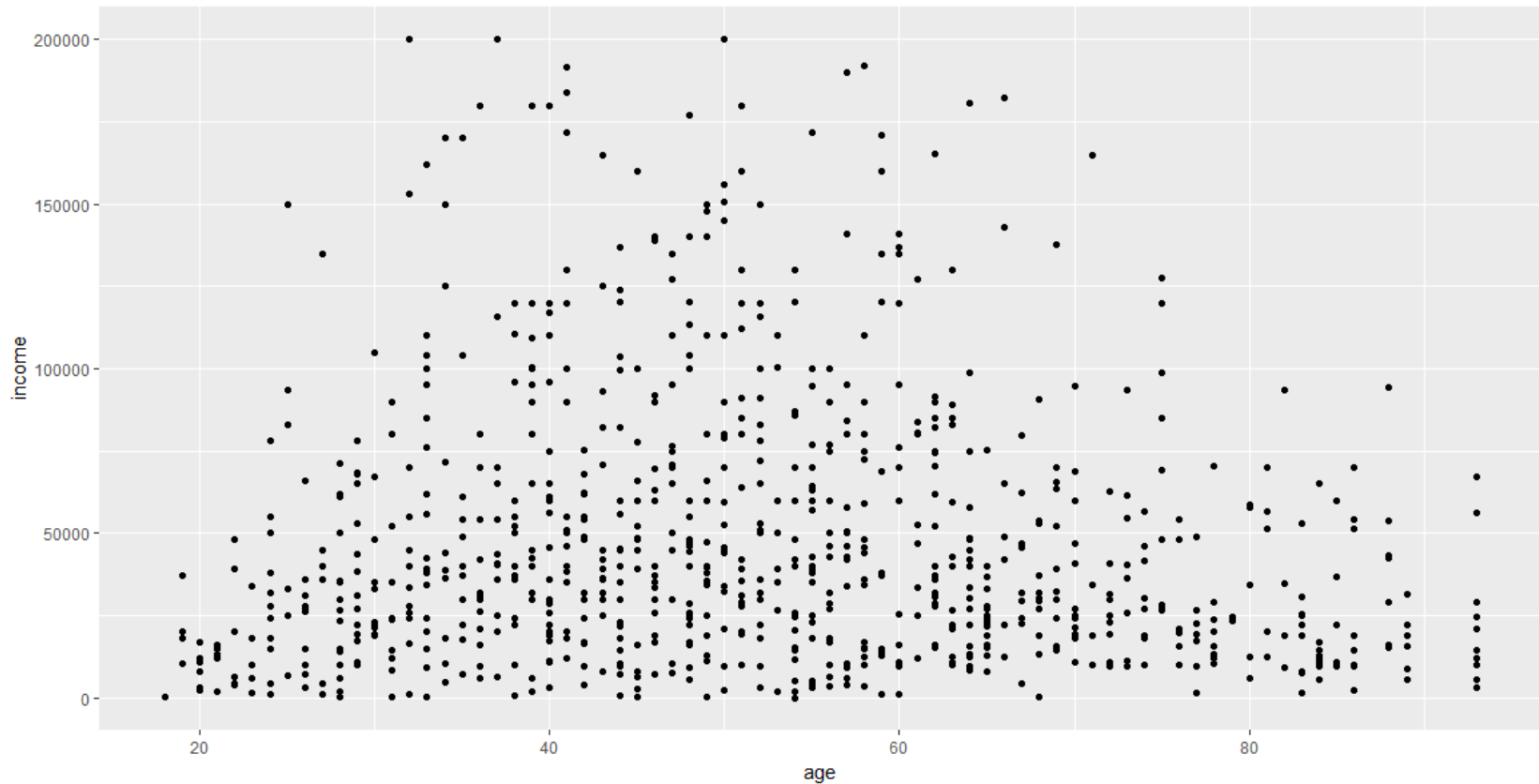


Gráfico de dispersión con linearización (geom_point)

- ¿Existirá una relación entre edad e ingreso?

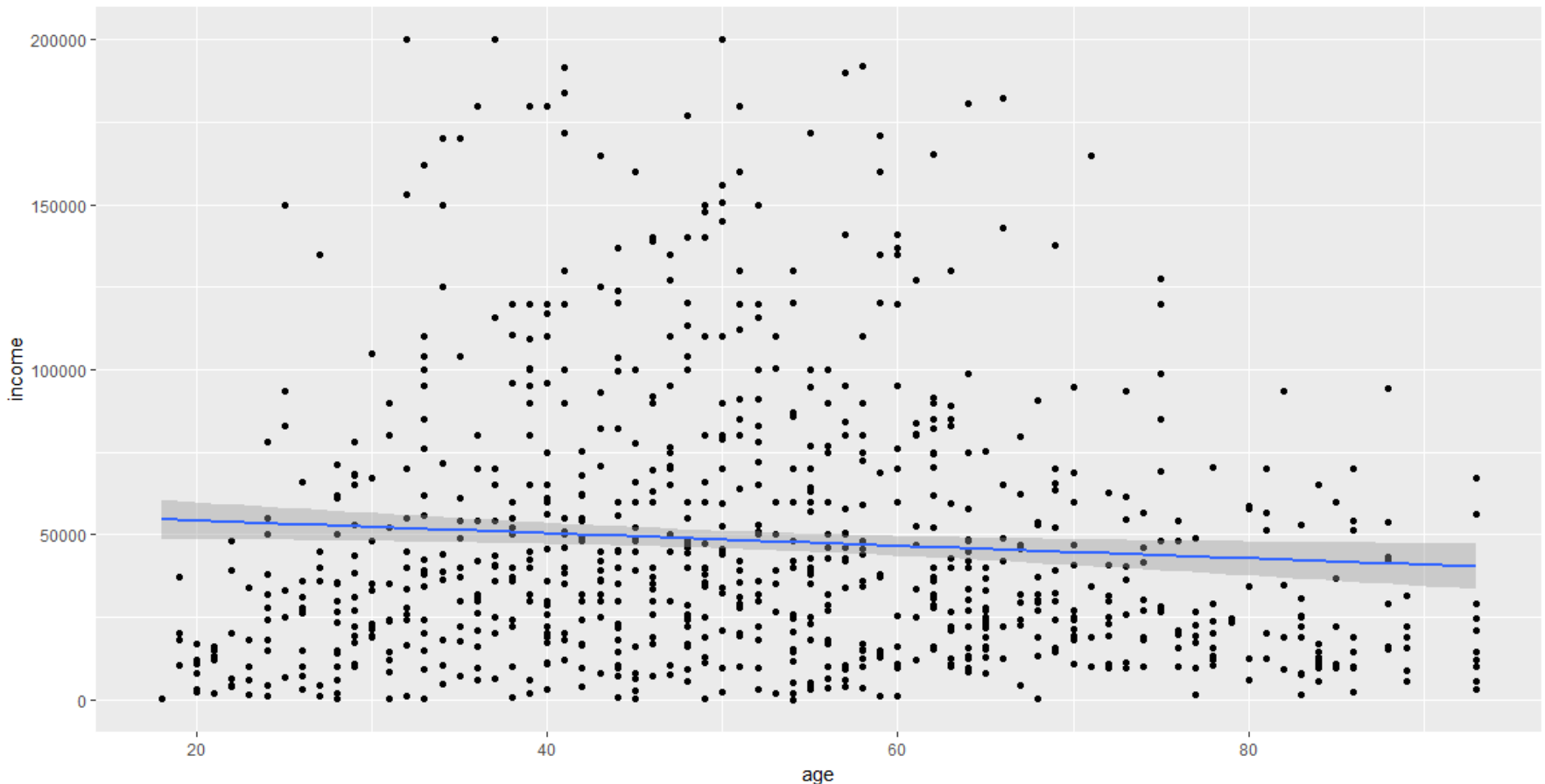
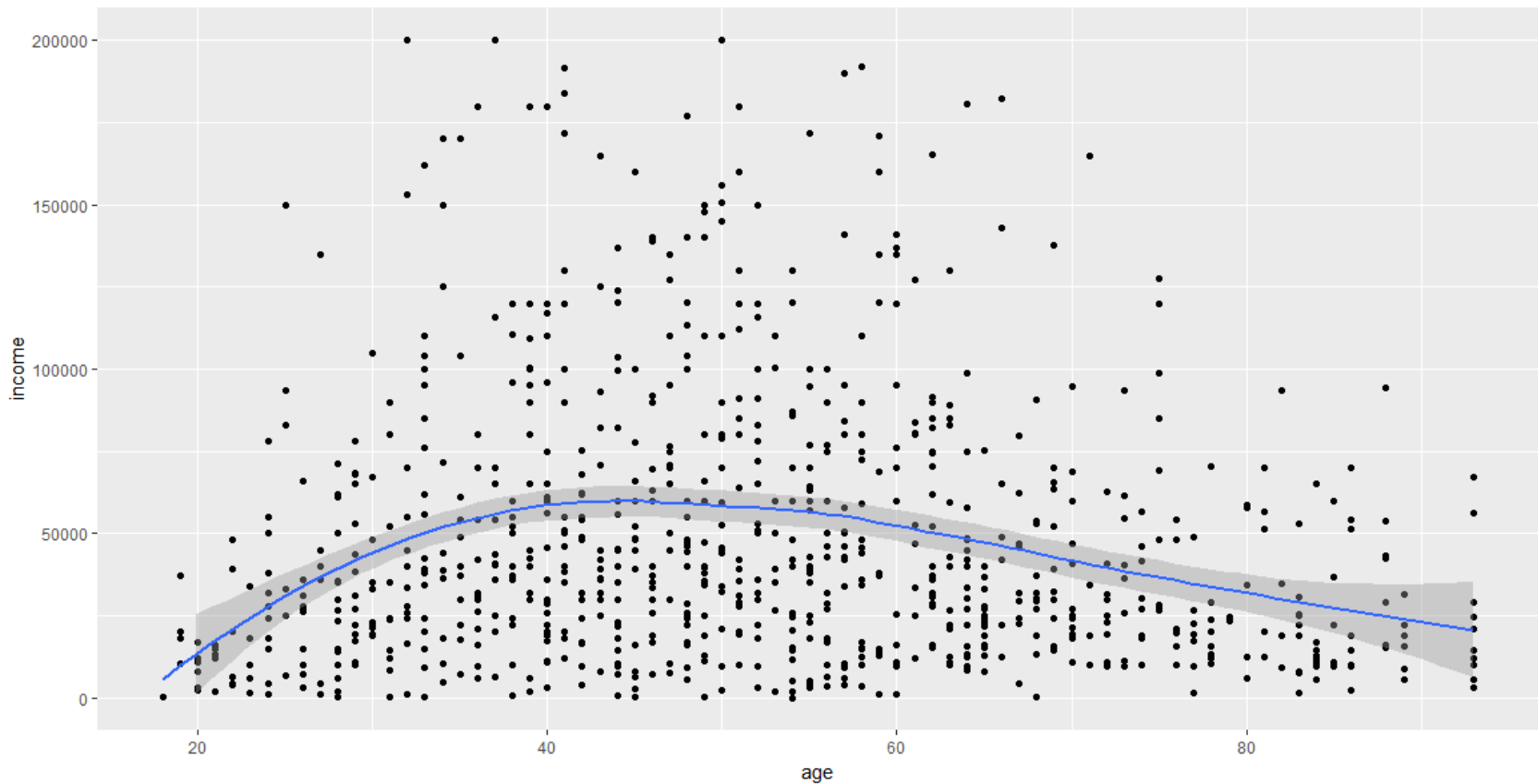
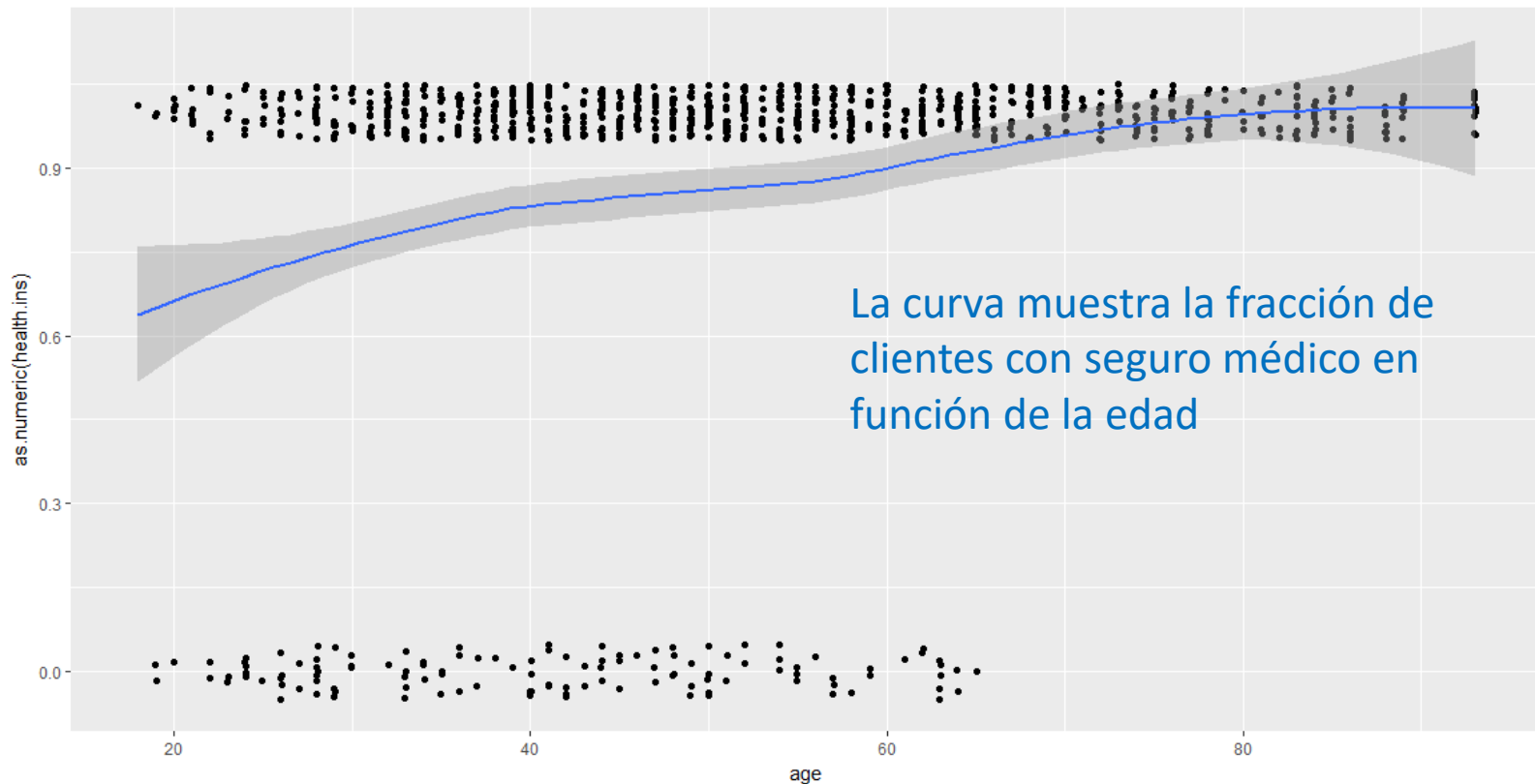


Gráfico de dispersión con curva de suavidad (geom_point)

- ¿Existirá una relación entre edad e ingreso?



¿Existe alguna relación entre el estado civil y el tener o no un seguro de salud?



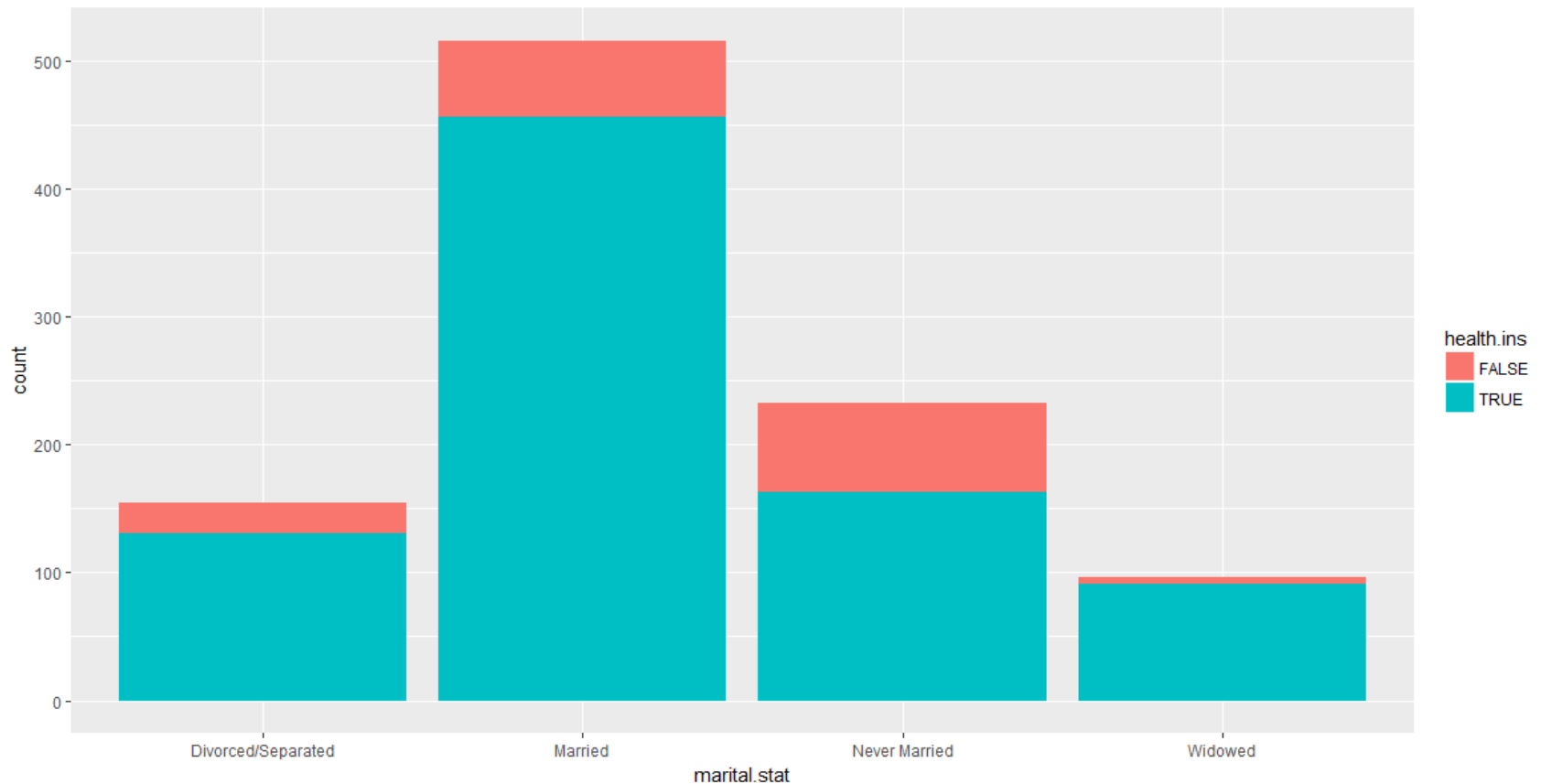
Correlación de variables categóricas

- Tablas cruzadas

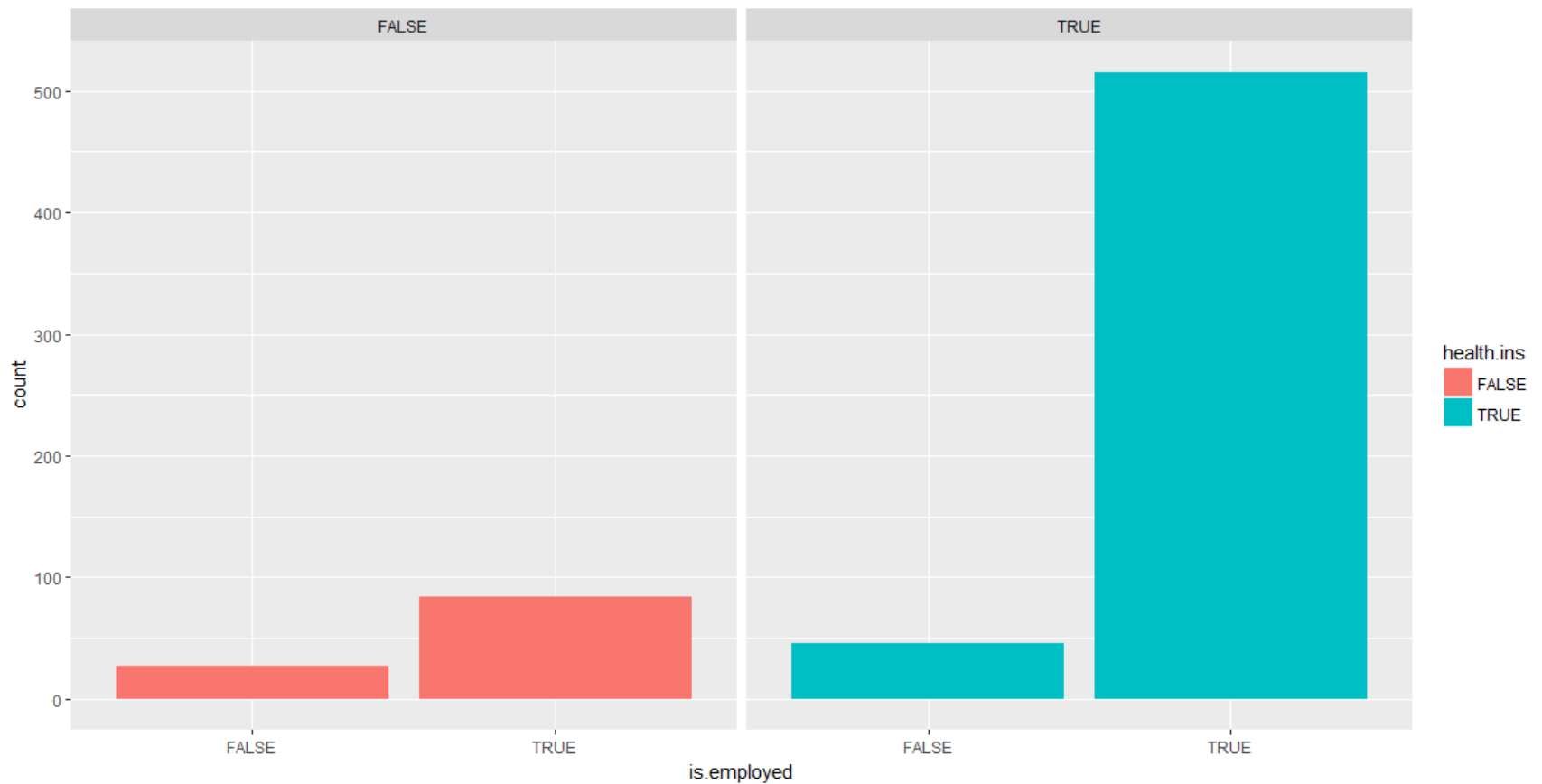
```
table(custdata$sex, custdata$health.ins)
```

	FALSE	TRUE
F	62	378
M	97	463

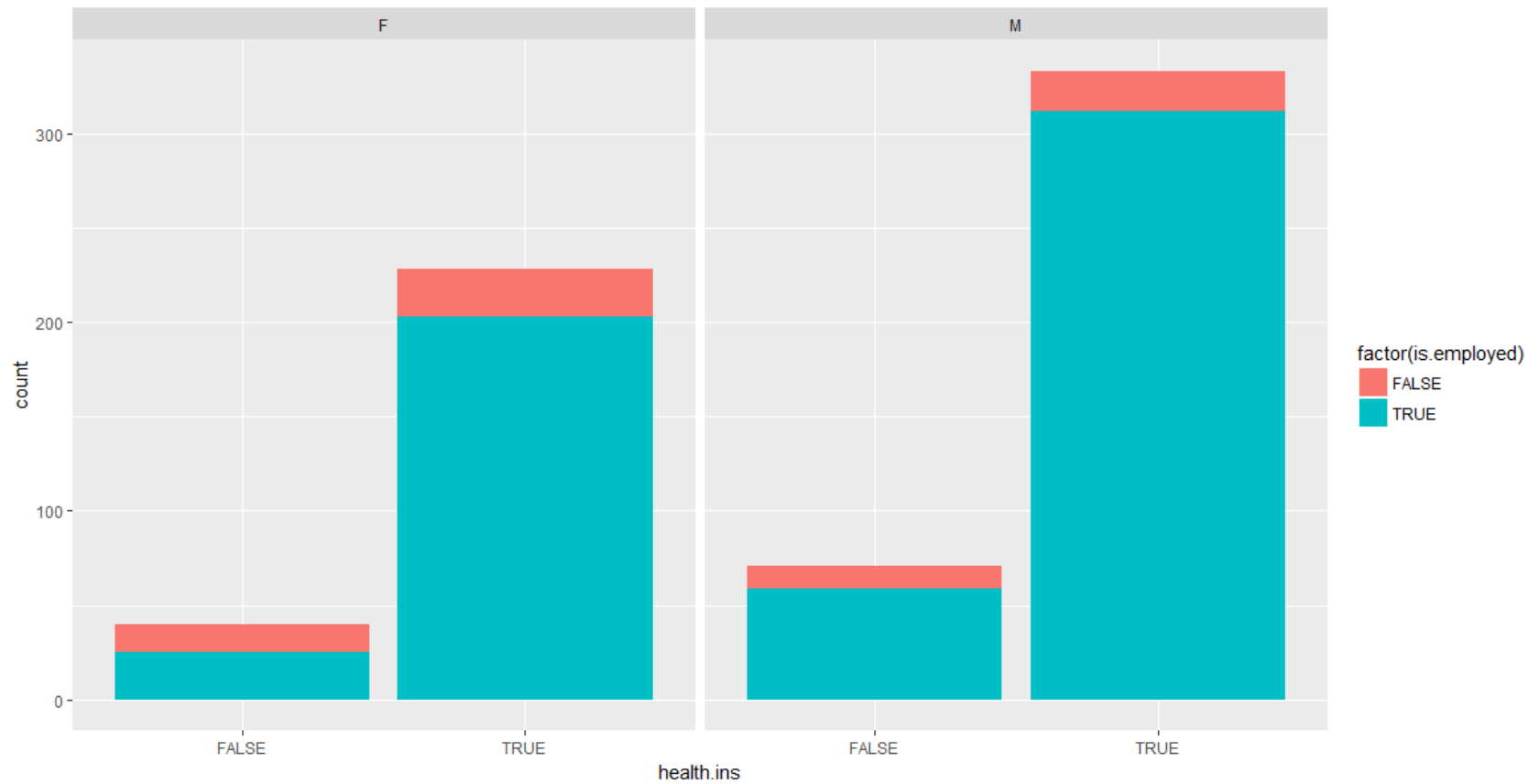
Gráficos de barras para examinar la correlación en variables categóricas



Gráficos de barras para examinar la correlación en variables categóricas



Más de análisis multivariado



Más de análisis multivariado

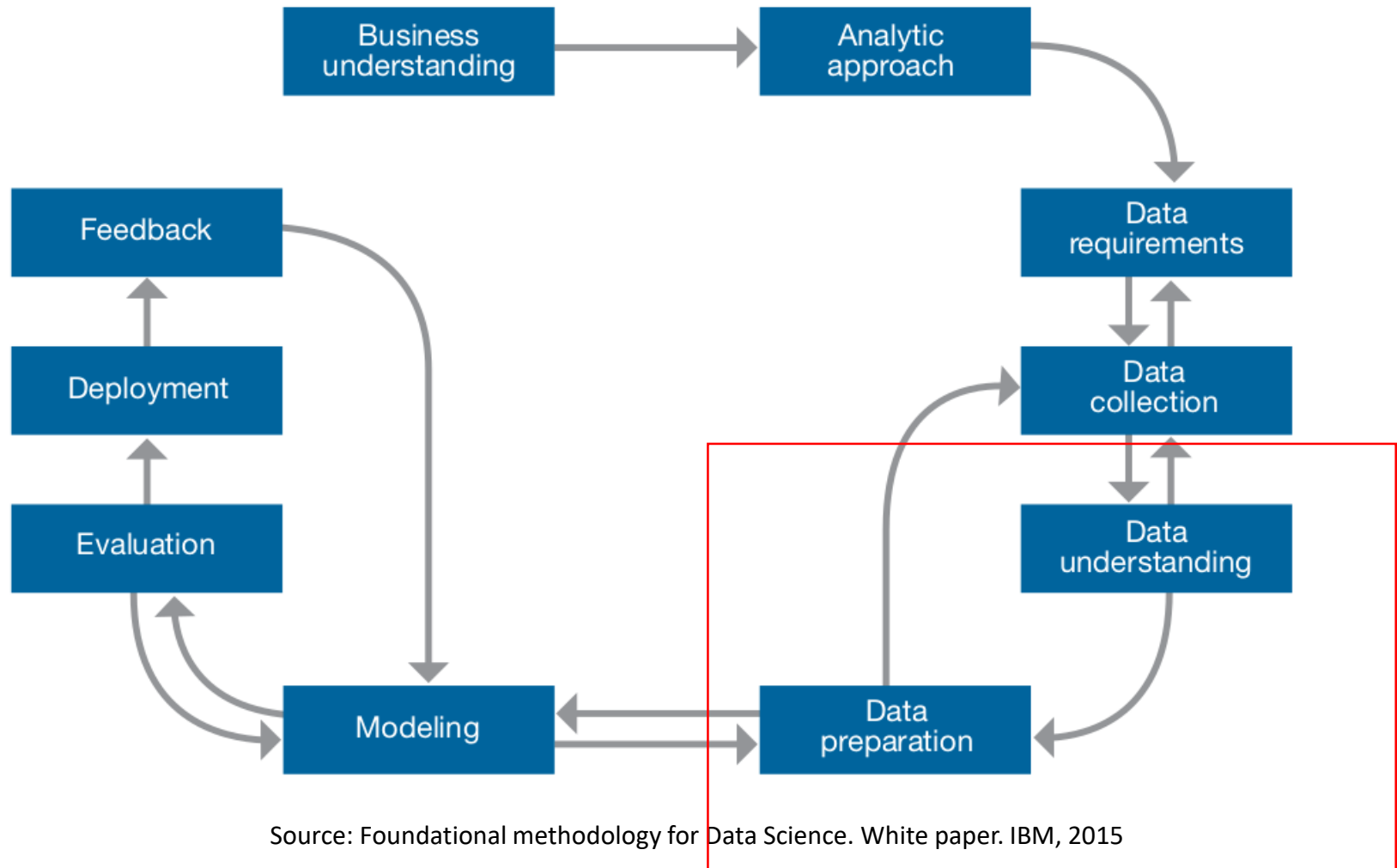


Retomemos la pregunta de analítica

- ¿Qué tipo de modelo podría servirnos para responder la pregunta del negocio?
- ¿Qué variables deberían usarse para el modelo?
- ¿Por qué fue importante realizar el análisis exploratorio?
- ¿Qué problemas identificamos en los datos?

Introducción a la Limpieza y Preparación de los Datos

Ciclo de vida de la analítica



Source: Foundational methodology for Data Science. White paper. IBM, 2015

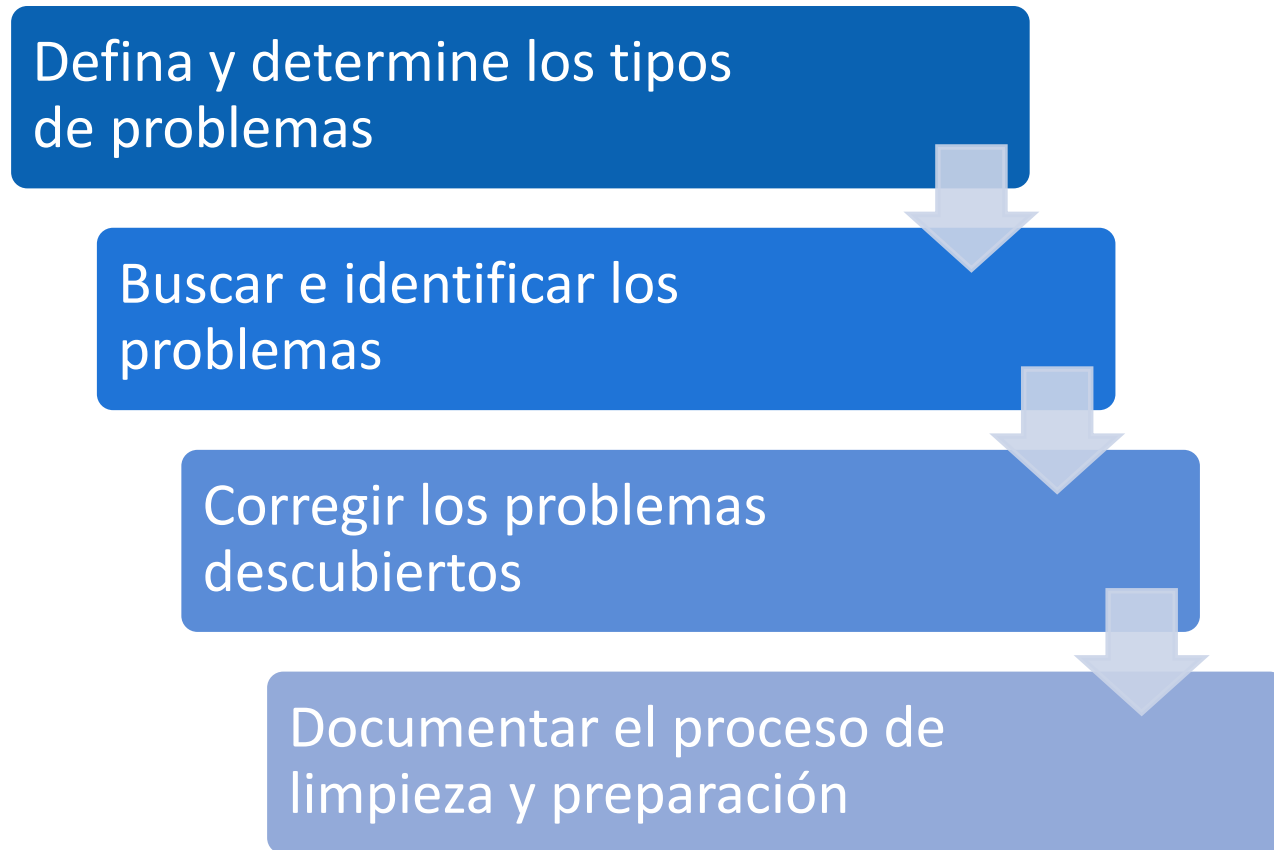
Datos brutos vs. datos preparados

- Datos brutos (raw data)
 - Son la fuente original de los datos
 - Difíciles de analizar
 - Usualmente se procesan solo una vez, para prepararlos
- Datos preparados (tidy data)
 - Listos para ser analizados
 - Su preparación incluye: limpieza, transformación, fusiones (merging), extracción de subconjuntos (subsetting)
 - **¡Todos los pasos ejecutados en la preparación deben documentarse para que sean repetibles!**

Más sobre los datos brutos

- Los datos brutos están en el formato adecuado si:
 - No se ha ejecutado ningún software sobre los datos
 - No ha habido manipulación de los datos
 - No se han eliminado datos
 - No se han resumido los datos

Proceso de limpieza y preparación



¡Consume la mayor parte del tiempo del proyecto!

Principales fuentes de problemas

Formato de las
variables no coincide
con el tipo de
variable

Observaciones
duplicadas
(coincidencia exacta)

Valores perdidos

Errores de digitación

Principales fuentes de problemas

Valores
inconsistentes

Valores fuera del
rango o inválidos

Valores sin referencia
en el diccionario de
variables

Separación de valores
de un campo en
varios campos

Exploremos un dataset bruto

Encuesta Americana de vivienda (American Housing Survey)

1. Descargue el dataset de las hipotecas de vivienda (Encuesta Americana de Vivienda) que está disponible a través del Moodle (Tema 1)
2. Explore el archivo descargado usando Excel (el separador es el ;)
 - ¿Qué problemas identifica usted en este dataset?
 - ¿Qué información representa la variable BANK?

Diccionario de variables (Cookbook)

- Fundamental para la limpieza, procesamiento y análisis de un conjunto de datos
- Describe el dataset
 - Variables, tipos
 - # de variables, # de observaciones
 - Entre otra información
- Insumo crítico en proyectos de analítica (Colombia: las empresas no suelen tenerlo)
- [Descargue el diccionario de variables de la encuesta Americana de vivienda \(disponible en Moodle\)](#)

Exploremos el dataset

- ¿Qué información representa la variable “ADDTNS”?
 - ¿Qué tipo de dato es, según el cookbook? ¿según Excel?
- ¿Qué información representa la variable “ADJPM”?
 - ¿Qué tipo de dato es, según el cookbook? ¿según Excel?

Documentación del proceso de limpieza y preparación

- Documentación paso a paso
- Debe incluir el código que se haya implementado
- Debe presentar todo el detalle para que el proceso sea repetible
- [Ejemplo](#)