Data Analytics (Statistics)
with Tableau Public

By: Iqbal Hanif

Profile

Institut Pertanian Bogor (2011 - 2015)

- S1 Statistika, Minor: Ekonomi & Studi Pembangunan
- Teaching Assistant: Metode Statistika & Perancangan Percobaan (2013-2014)
- Intern di SAS Institute (2014)
- Research Assistant di Bank Indonesia (2015)

Telkom Indonesia (2016 – now)

- Trainee - GPTP IV (2016)
- Officer 3 Data Scientist (2017)
- Officer 2 Data Scientist / Big Data Analytics (2020)

Outline

1. **Tableau & Statistics**
   - Tableau Overview
   - Statistics Overview

2. **Descriptive Statistics**
   - Usecase: Segmentation - HVC
   - Usecase: Outlier Detection - Overspec & Underspec

3. **Inferential Statistics**
   - Usecase: A/B Testing - Usability Testing
   - Usecase: Regression - Marketing Cost Analysis

DATA SCIENCE INDONESIA

Tableau & Statistics

Overview

Tableau

Tableau is a **visual analytics platform** transforming the way we use data to **solve problems**—empowering people and organizations to make the most of their data (tableau.com)

Tableau is business intelligence software that allows anyone to **connect to data** in a few clicks, then **visualize** and create interactive, sharable **dashboards** (softwareconnect.com)

Tableau Products

DATA SCIENCE INDONESIA

Commercial

Tableau Desktop
Create
FOR ANYONE

Tableau Server
Share & Create
FOR ORGANIZATIONS

Tableau Online
Share & Create
FOR ORGANIZATIONS

- Personal/Professional Edition
- Student Edition (max 1 year)

Free

Tableau Reader
Share
FOR ANYONE

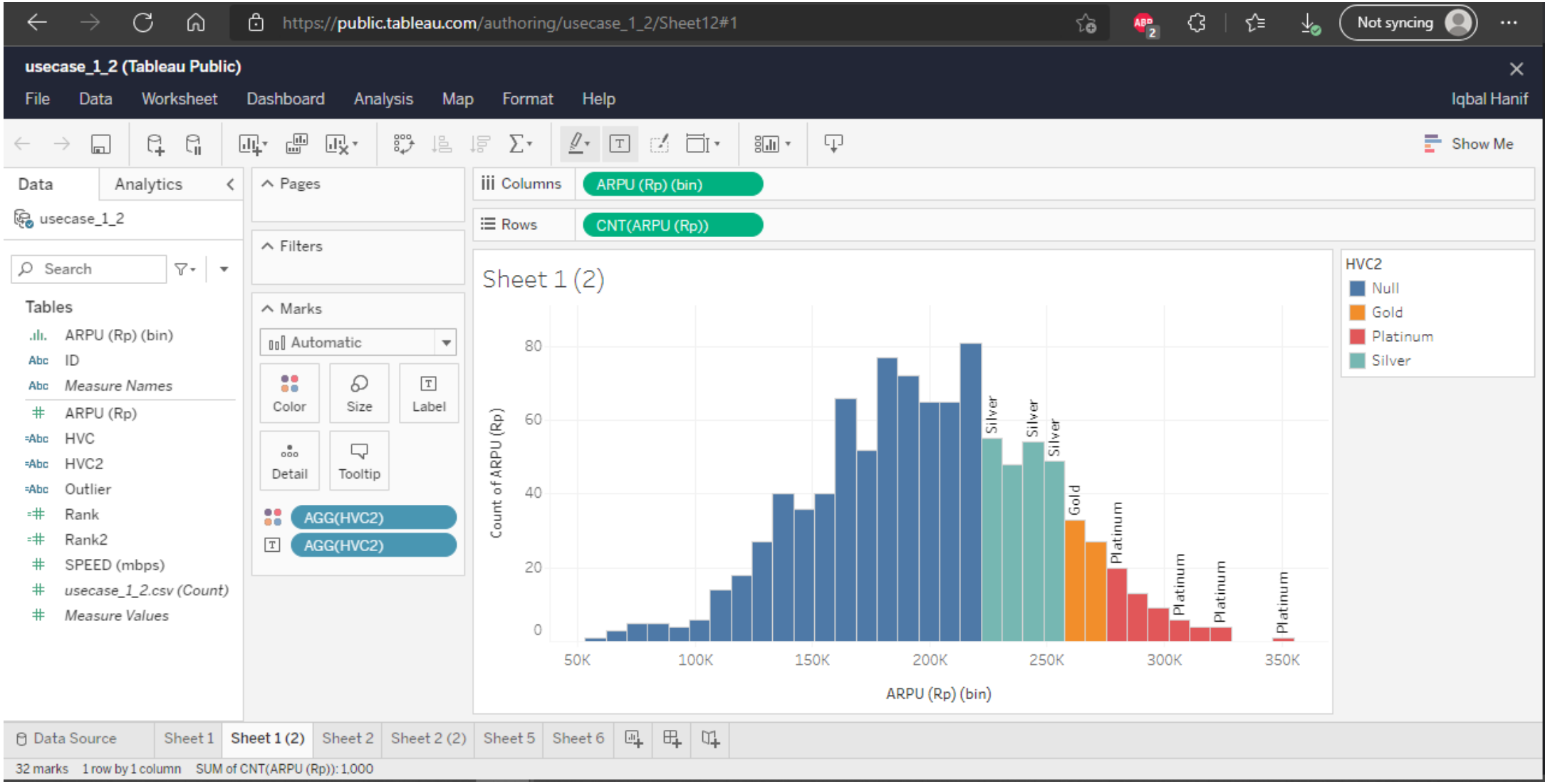
Tableau Public
Share & Create
FOR INDIVIDUALS

- Tableau Public have 2 version (dekstop & online)

Others:
- Tableau Prep
- Tableau Mobile

Tableau Public (Online)

DATA SCIENCE INDONESIA



Statistics Overview

DATA SCIENCE INDONESIA

Descriptive Statistics

Helps us to show, summarize and analyze **how our collected data characteristic.** Descriptive statistic enable us to represent raw data in a meaningful way

Inferential Statistics

Helps us to predict, generalize fact from sampling data to describe and **give a picture how the population data characteristic.**

Level of Measurement

Nominal

Categorical value which every value could be distinguished. Different value indicates different object

GENDER

COUNTRIES

MARITAL STATUS

Ordinal

The categorical value has meaningful order and can be sorted

RANK

ASSIGNED FREQUENCY

Interval

Numeric scales in which we know both the order and the exact differences between the values, they don't have a "true zero."

TEMPERATURE

EXAM SCORE

Ratio

The exact value between units, AND they also have an absolute zero.

HEIGHT

WEIGHT

Descriptive Statistics

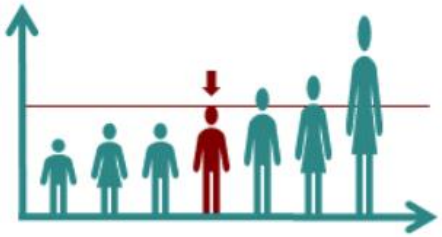
Ukuran Tendensi Sentral	Ukuran Letak	Ukuran Penyebaran	Bentuk Sebaran
• Mean • Median • Mode	• Quartil • Desil • Persentil	• Variance • Range • Interquartile Range (IQR)	• Modality • Skewness

Descriptive Statistics Overview

Mean

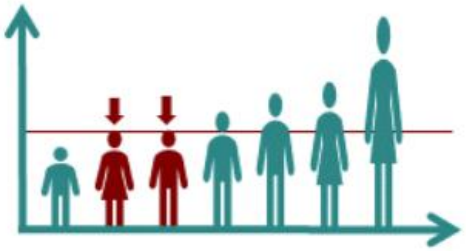
Sum of all values divided by the number of all values.

Median



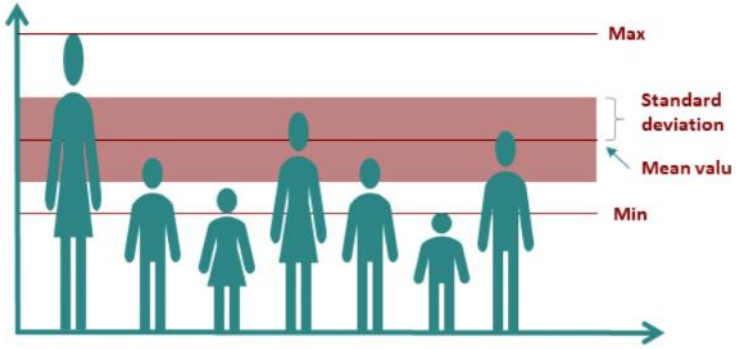
Above the value and below the value are the same number of cases. It halves the distribution.

Mode



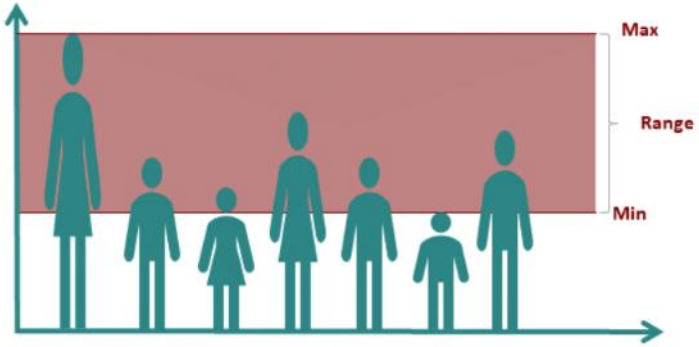
Occurs most frequently in a distribution

Standard Deviation & Variance



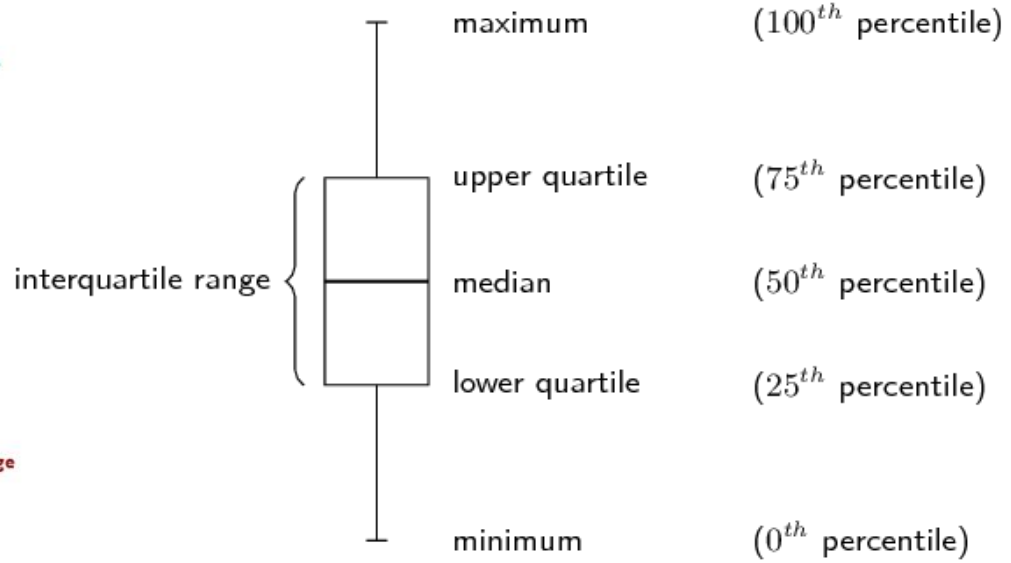
measure of how much a sample fluctuates around a mean value.

Range

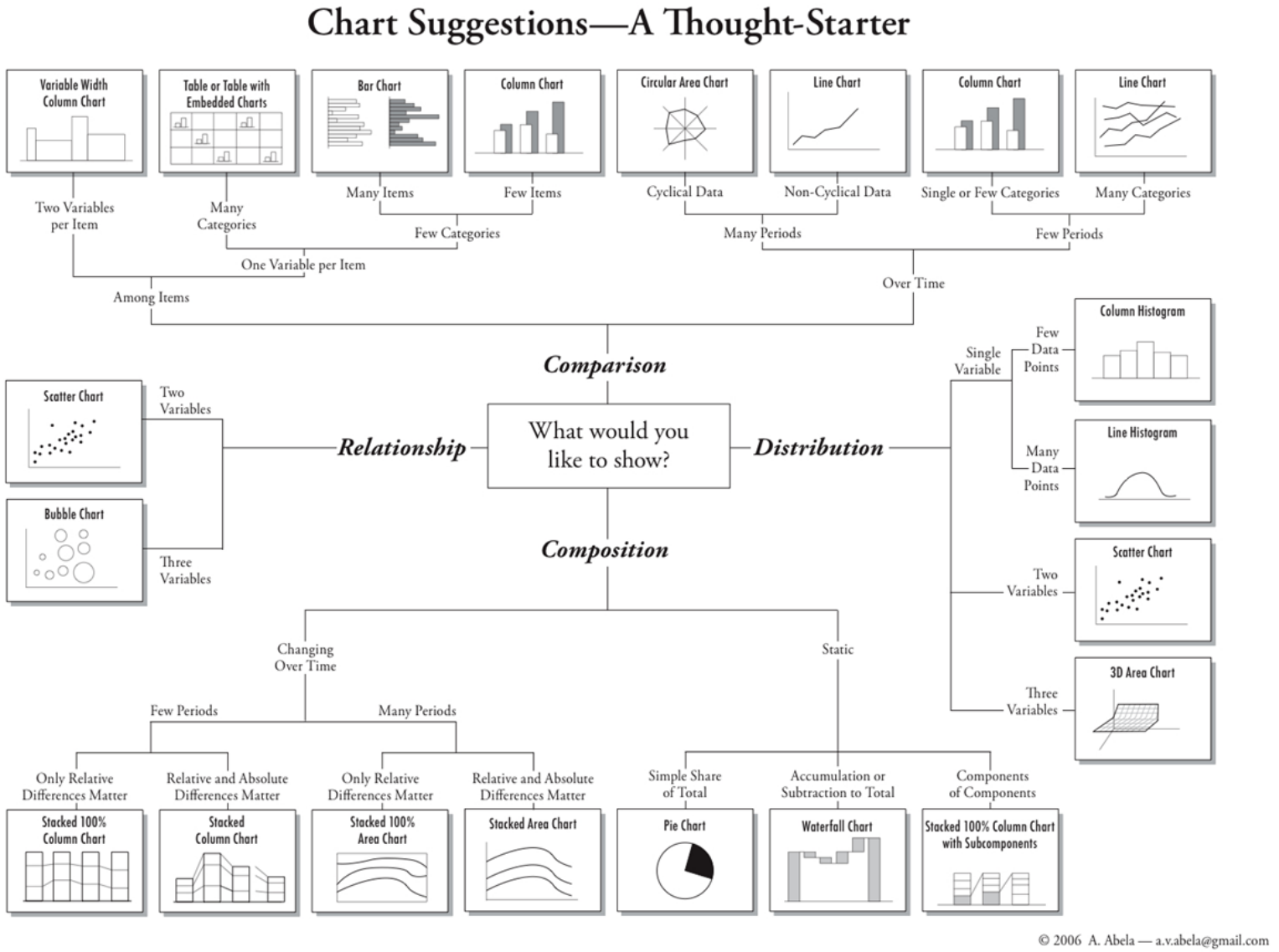


the distance between the highest and the lowest value

Five Number Summary



Descriptive Statistics Overview



Inferential Statistics

DATA SCIENCE INDONESIA

Regression Analysis	Correlation Analysis	Confidence Interval	Hypothesis Testing
• Linear Regression • Logistic Regression .	• Pearson Correlation • Spearman Rank Correlation	• Mean CI • Proportion CI	• T-test • ANOVA • Chi Square

Inferential Statistics Overview

DATA SCIENCE INDONESIA

Predictor / Covariate / Independent Variable		Criterion / Measure / Dependent Variable: Categorical — 1 Variable, 2 Categories	Categorical — 1 Variable, >2 Categories	Continuous — 1 Variable	Continuous — >1 Variable
Categorical	1 Variable 2 Categories Between-subjects	χ^2 Test (Crosstabs → Statistics → ☑ Chi-square)		**Independent *t* Test** (Compare Means → Independent-Samples)	
	1 Variable 2 Categories Within-subjects			**Paired *t* Test** (Compare Means → Paired Samples)	
	1 Variable >2 Categories Between-subjects			**One-Way ANOVA** (Compare Means → One-way ANOVA)	**One-Way MANOVA** (General Linear Model → Multivariate → Add Dependent Variables)
	1 Variable >2 Categories Within-subjects			**Repeated Measures ANOVA** (General Linear Model → Repeated Measures → Add Within-Sbj Factors)	**Repeated Measures MANOVA** (Repeated Measures ANOVA → Add Measures)
	>1 Variable All Categorical Between-subjects	**Binomial Logistic Regression with Categorical Predictors** (Regression → Binary Logistic → Categorical Covariates)	**Multinomial Logistic Regression** (Regression → Multinomial Logistic)	**Factorial ANOVA** (General Linear Model → Univariate → Add Fixed Factors)	**Factorial MANOVA** (One-Way MANOVA → Add Fixed Factors)
	>1 Variable All Categorical Mixed Within- & Between-subjects			**Mixed-Design ANOVA** (Repeated Measures ANOVA → Add Between-Sbj Factors)	**Mixed-Design MANOVA** (Mixed-Design ANOVA → Add Measures)
	>1 Variable Mixed Categorical & Continuous			**One-Way ANCOVA** (One-Way ANOVA → Add Covariates)	**One-Way MANCOVA** (One-Way MANOVA → Add Covariates)
Continuous	1 Variable	**Binomial Logistic Regression** (Regression → Binary Logistic)		**Simple Linear Regression** (Regression → Linear)	**Multivariate Linear Regression** (General Linear Model → Multivariate → Add Dependent Variables → No Fixed Factors → Add Covariates)
	>1 Variable			**Multiple Linear Regression** (Regression → Linear)	

DATA SCIENCE INDONESIA

Descriptive Statistics

Usecase

Customer Segmentation

- Membagi data pelanggan menjadi beberapa kelompok berdasarkan kriteria tertentu.
- Tujuannya adalah untuk membedakan treatment/perlakuan yang akan diberikan antara satu kelompok dan kelompok lainnya demi mencapai tujuan bisnis.
- Salah satu contoh penerapan segmentasi adalah High Value Customer (HVC), yaitu segmen khusus untuk pelanggan yang berpotensi memberikan revenue yang lebih tinggi dibanding dengan pelanggan-pelanggan lainnya.

Case 1: Customer Segmentation - High Value Customer

- Anda bersemangat menyongsong hari pertama bekerja sebagai data analis di perusahaan operator seluler ABC.
- Di tugas pertamanya, Anda tergabung dalam tim Customer Relationship Management (CRM) yang sedang berupaya membuat segmen HVC untuk pelanggan produk baru yang diluncurkan beberapa bulan yang lalu.
- Masing-masing anggota tim diminta untuk menganalisa data yang tersedia serta membuat draft/rancangan metode terbaik yang bisa digunakan untuk menentukan segmen HVC.
- Dari data yang tersedia, Anda tertarik menggunakan data ARPU (average revenue per user) yaitu rata2 saldo pulsa (Rp) yang dibeli pelanggan per bulan.

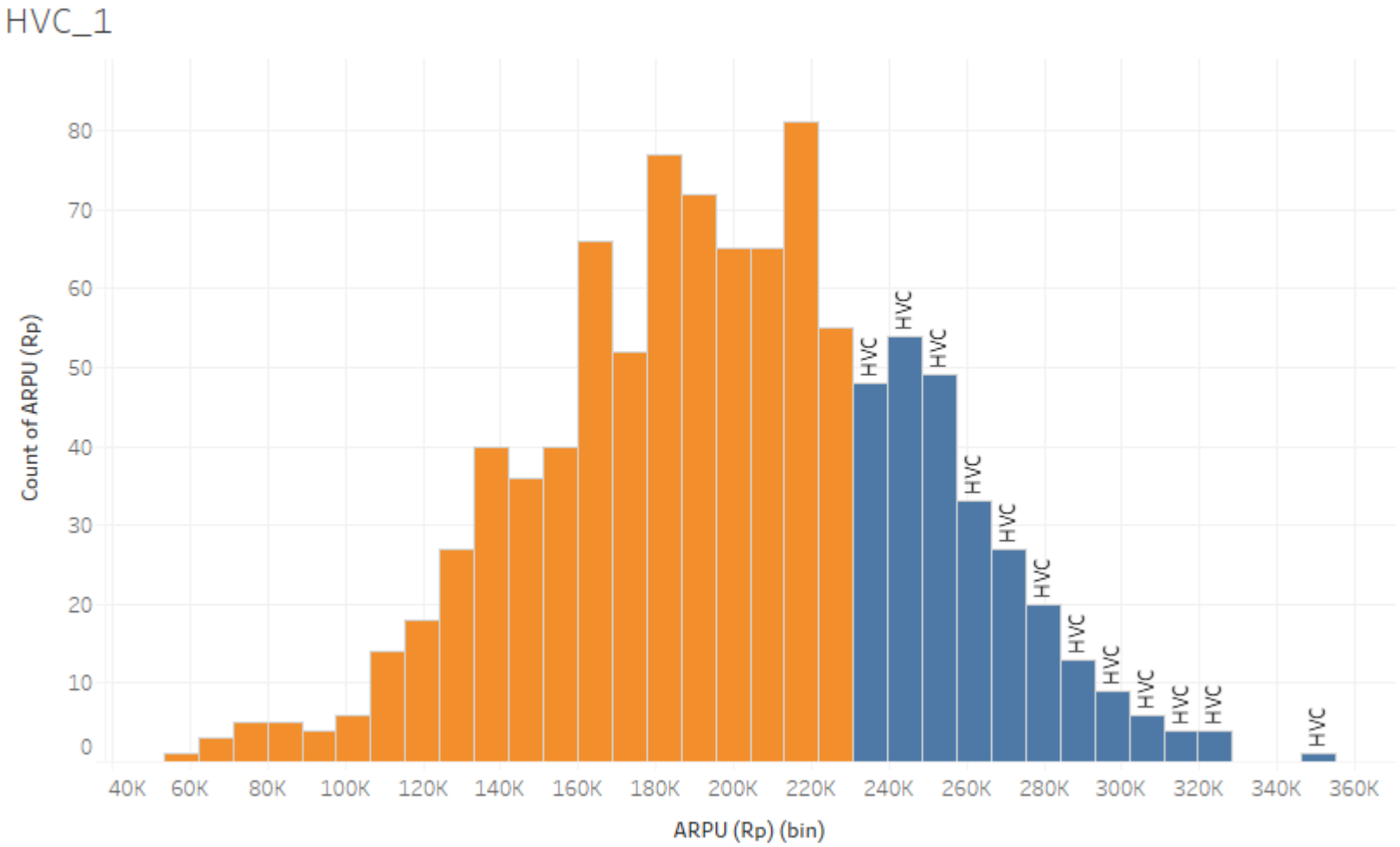
Case 1: Customer Segmentation - High Value Customer

- Dengan menggunakan statistika deskriptif, metode perhitungan apakah yang cocok dalam menentukan titik pembeda antara segmen HVC dan non HVC?

Ukuran Letak

a. Quartil
b. Desil
c. Persentil

Case 1: Customer Segmentation - High Value Customer



Case 1: Customer Segmentation - High Value Customer

Quartile Formula

$$\text{Lower Quartile (Q1)} = (N+1) \times \frac{1}{4}$$

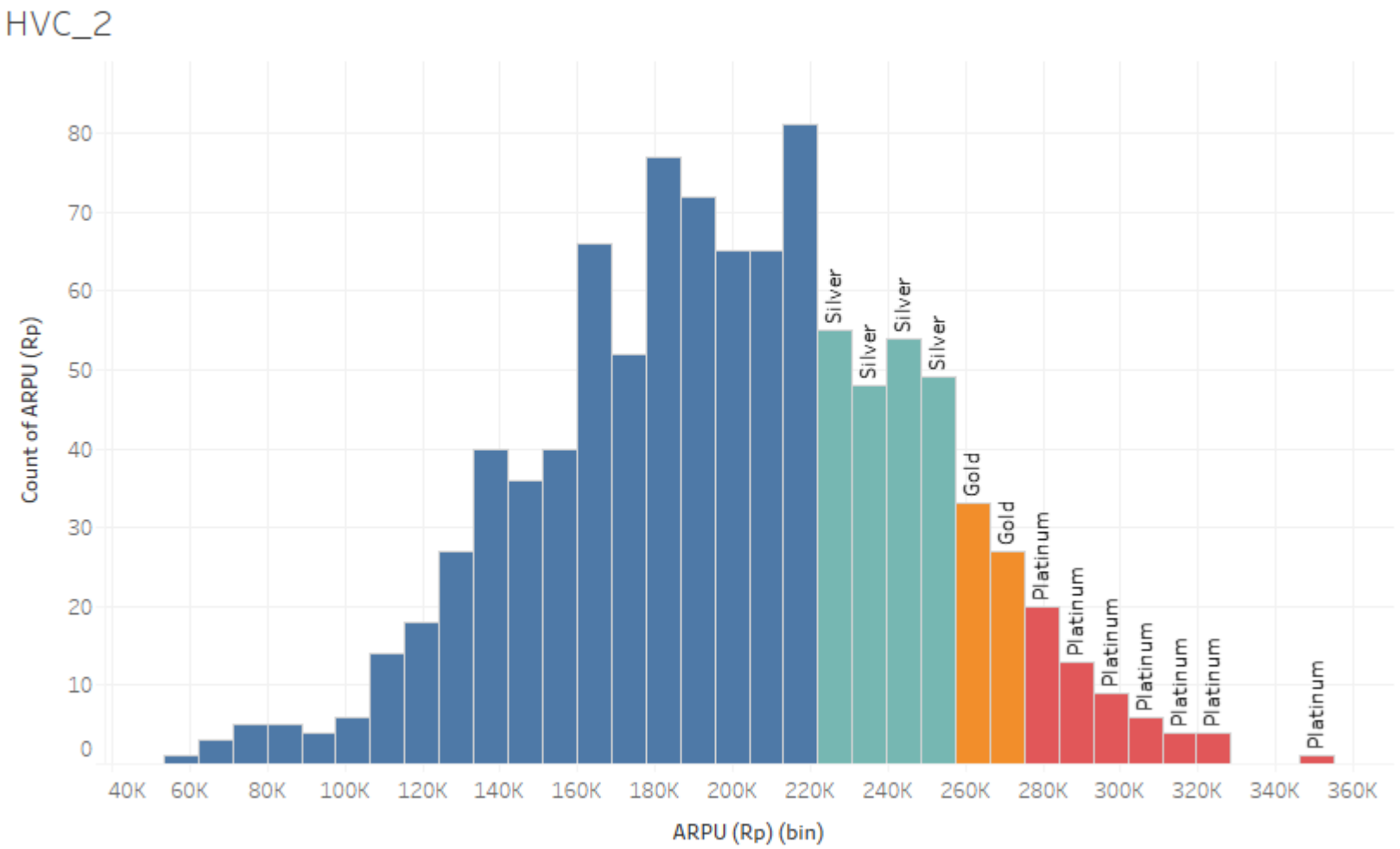
$$\text{Middle Quartile (Q2)} = (N+1) \times \frac{2}{4}$$

$$\text{Upper Quartile (Q3)} = (N+1) \times \frac{3}{4}$$

Percentile Rank Formula

$$R = \frac{P}{100\,(N+1)}$$

Case 1: Customer Segmentation - High Value Customer



Case 1: Customer Segmentation - High Value Customer

- Treatment apa yang akan anda berikan kepada HVC?

Treatment
a. Diskon untuk pembelian pulsa/potongan untuk pasca bayar b. Poin untuk ditukarkan voucher diskon/barang serta grand prize bulanan/tahunan c. Prioritas antrian untuk layanan gangguan atau CS d. Add on tambahan (Netflix, Disney+, Spotify, etc).

Outlier Detection

- Menemukan data yang bersifat pencilan (outlier), yaitu data yang memiliki karakteristik yang jauh berbeda dengan data-data lainnya.
- Tujuannya adalah untuk mendeteksi anomali yang terjadi dalam suatu proses bisnis, yang bisa saja merupakan indikasi dari suatu kejadian yang tidak diinginkan (fraud, human error, system error, dll.)
- Salah satu contoh penerapan outlier detection adalah deteksi Underspec & Overspec, yaitu kondisi dimana layanan yang digunakan pelanggan tidak sesuai dengan spesifikasi yang disepakati, bisa jauh lebih rendah (under specification) atau jauh lebih tinggi (over specification).

Case 2: Outlier Detection - Overspec & Underspec

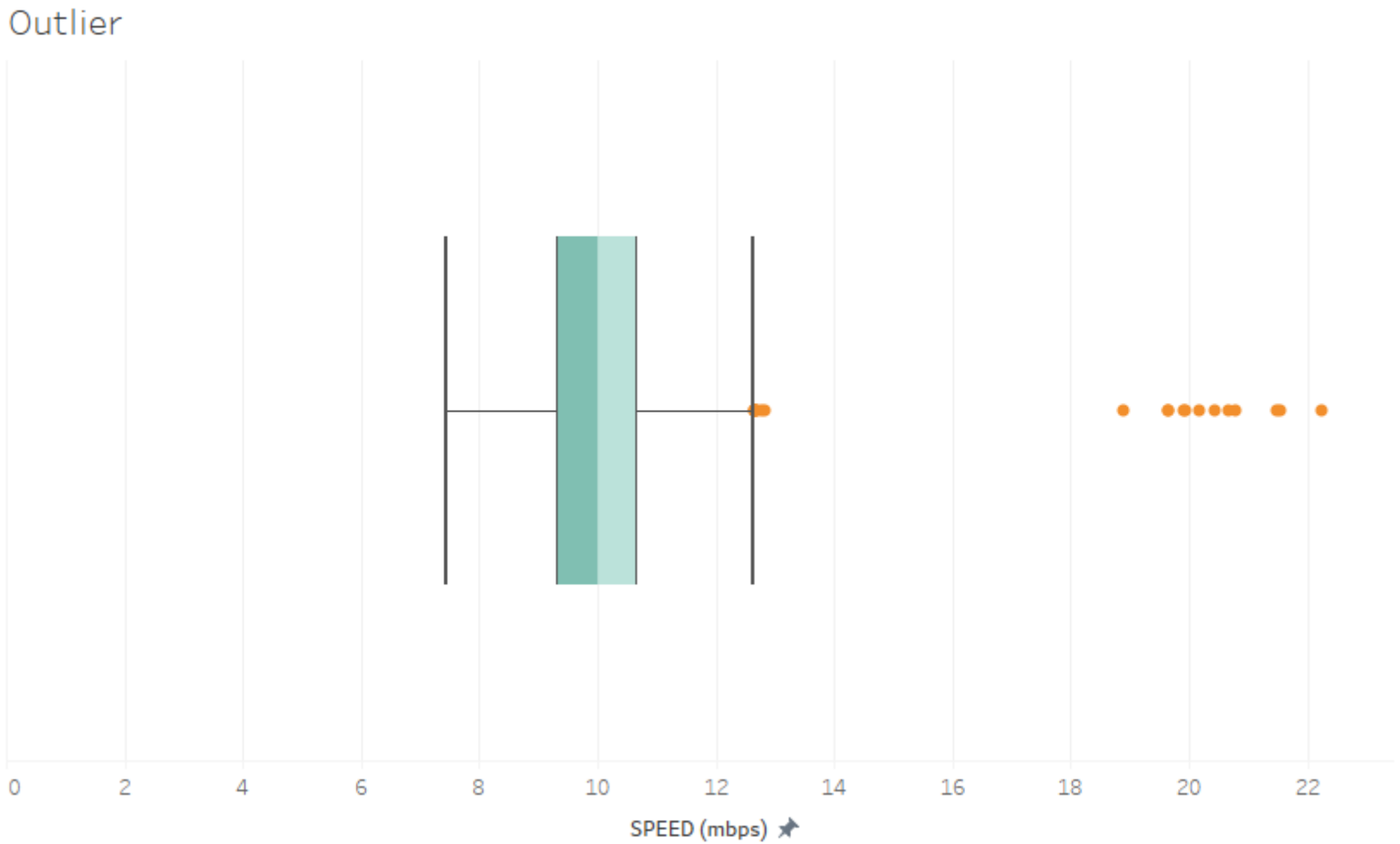
- Anda adalah seorang data analyst perusahaan provider internet yang dipindahtugaskan dari departemen sales ke departemen IT & network.
- Di tempat baru, Anda bertanya mengenai apa problem yang selama ini dialami oleh tim network, kemudian Anda menemukan bahwa teknisi sering mengeluh karena ada penggunaan internet yang tidak wajar sehingga berisiko menimbulkan gangguan pada pelanggan.
- Anda mulai mencoba mengidentifikasi data yang bisa digunakan untuk mendeteksi keanehan tersebut, Tim teknisi memberikan akses data kecepatan internet (speed) kepada Anda. Untuk riset awal, Anda ingin meneliti data speed pada produk internet dengan kecepatan 10 mbps.

Case 2: Outlier Detection - Overspec & Underspec

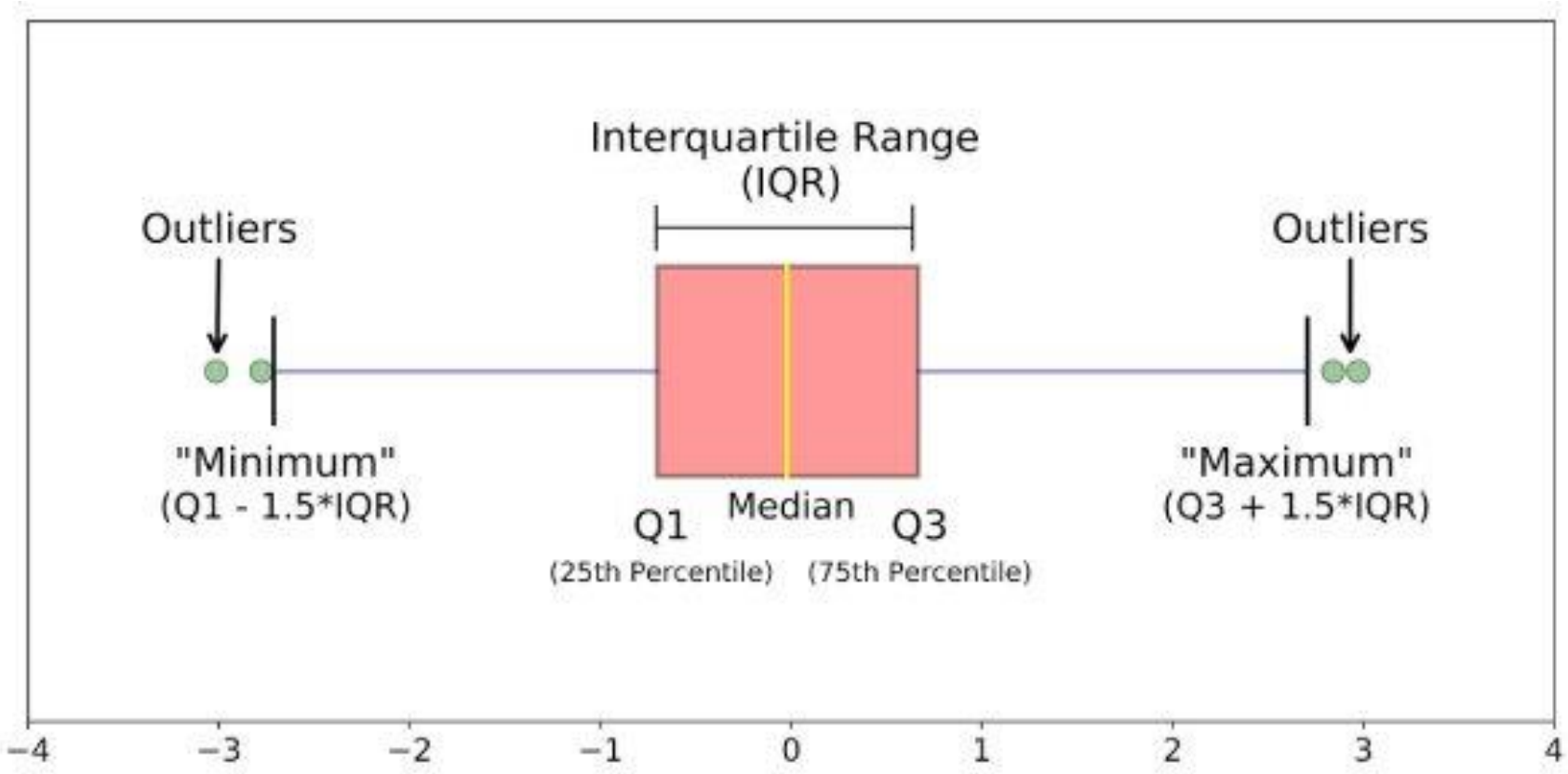
- Dengan menggunakan statistika deskriptif, metode perhitungan apakah yang cocok dalam menentukan data overspec atau underspec (outlier)?

Ukuran Penyebaran
a. Variance/Standard Deviation b. Range c. Interquartile Range (IQR)

Case 2: Outlier Detection - Overspec & Underspec



Case 2: Outlier Detection - Overspec & Underspec



Case 2: Outlier Detection - Overspec & Underspec

- Langkah apa yang akan anda lakukan kepada Overspec?

Next Action
a. Pemeriksaan teknis langsung perangkat-perangkat di lapangan b. Pemberhentian layanan pelanggan yang terindikasi fraud c. Sidak pegawai lapangan yang terindikasi lalai (human error) atau membantu fraud. d. Memeriksa sistem/database terkait kemungkinan salah pencatatan (system error)

How About Machine Learning?

(+) Statistics Descriptive

- Variabel/feature yang digunakan sedikit.
- Lebih mudah dipahami/diinterpretasikan ke user/customers
- Kebutuhan komputasi lebih ringan

(+) Machine Learning

- Variabel/feature yang digunakan banyak.
- Pattern yang diinginkan belum diketahui/ditemukan.

DATA SCIENCE INDONESIA

Inferential Statistics

Usecase

A/B Testing

DATA SCIENCE INDONESIA

- Uji coba yang dilakukan untuk membandingkan dua atau lebih produk atau metode dengan melakukan uji statistik.
- Tujuannya adalah untuk menentukan metode atau produk mana yang terbaik dengan hanya melakukan uji coba dalam skala kecil (sampling), dan uji statistik membuat hasil pengujian menjadi lebih akurat dan dapat diterima oleh populasi ke depannya.
- Salah satu contoh penerapan A/B Testing adalah Usability testing, yaitu uji coba yang dilakukan terhadap sebuah aplikasi yang sedang di develop sebelum di luncurkan dan digunakan oleh user. Usability testing berguna untuk mengevaluasi kekurangan pada aplikasi dan memprediksi apakah aplikasi akan diterima oleh user atau tidak di masa mendatang.

Case 3: A/B Testing: Usability Testing

- Anda adalah seorang data analis di sebuah start up unicorn. Anda bertugas di tim produk yang baru dikembangkan yaitu jasa food delivery.
- Start up tersebut masih berupaya mengoptimalkan produk mereka hingga disukai oleh pengguna (user). Salah satu caranya adalah mengganti interface aplikasi mereka (UI) sehingga pelanggan semakin nyaman (UX).
- Sebelum merilisnya, tim produk akan melakukan usability testing untuk menguji interface baru aplikasi mereka. Salah satu metrik pengukuran yang digunakan adalah lama penggunaan aplikasi per hari (menit), dimana responden penelitian akan diukur waktu penggunaan dengan aplikasi lama (A), kemudian diukur waktu penggunaan dengan aplikasi baru (B).

Case 3: A/B Testing: Usability Testing

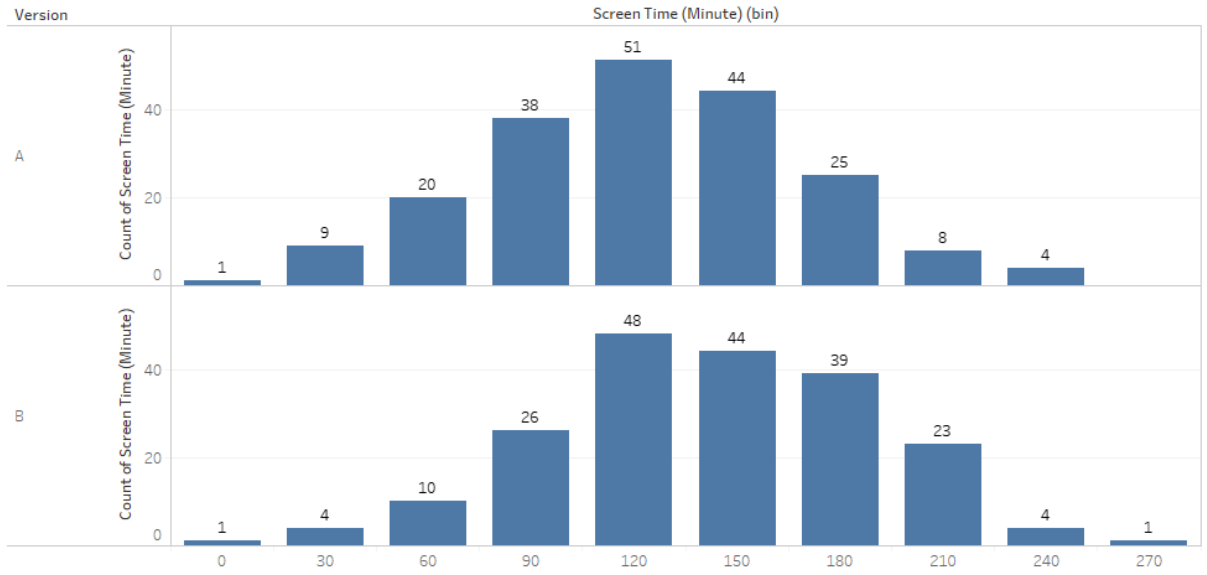
- Dengan menggunakan statistika inferensia, metode apakah yang cocok dalam menentukan apakah tampilan terbaru (B) lebih baik dibandingkan tampilan lama?

Hypothesis Testing

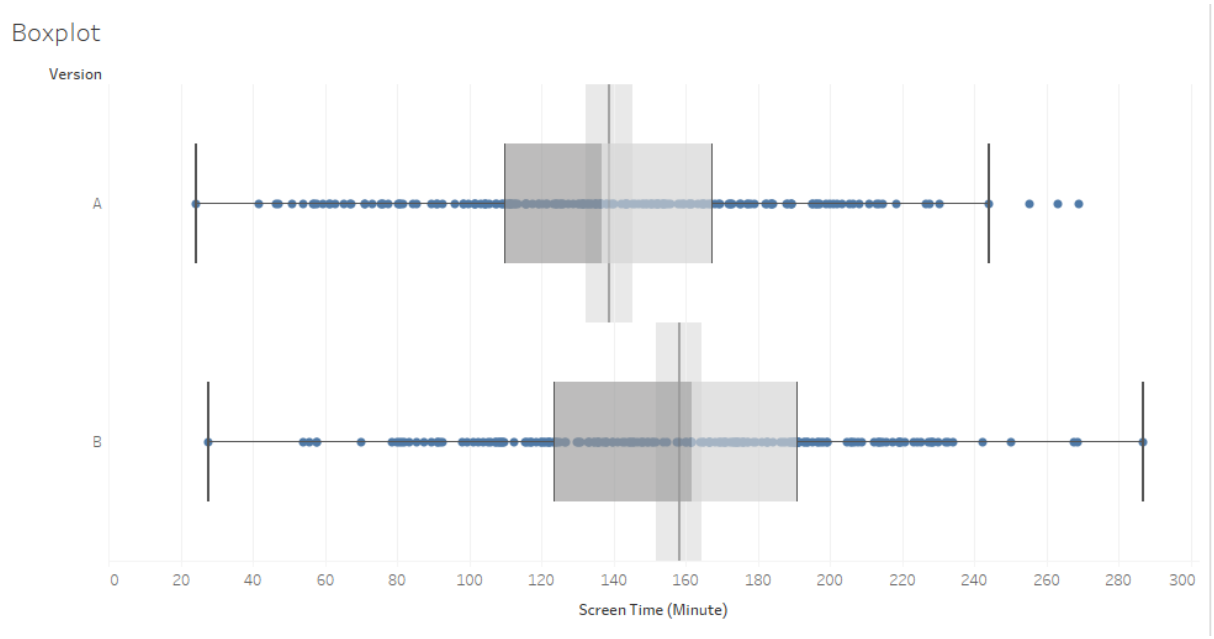
a. T-test
b. ANOVA
c. Chi Square

Case 3: A/B Testing: Usability Testing

Histogram



Boxplot



Case 3: A/B Testing: Usability Testing

Paired T-Test Summary

T-Statistics Numerator	19.40
T-Statistics Denumerator	4.94
T-Statistics	3.93
T-Statistic Table	1.96

Case 3: A/B Testing: Usability Testing

Our hypotheses:

$$H_o: \mu_D = 0$$

$$H_A: \mu_D \neq 0$$

Reject H0 if:

- $|t| > t(\alpha/2, n-1)$
- P-value < α

$$t = \frac{\Sigma d}{\sqrt{\frac{n(\Sigma d^2) - (\Sigma d)^2}{n-1}}}$$

where d: difference per paired value

n: number of samples

$$t = \frac{(\sum D)/N}{\sqrt{\frac{\sum D^2 - \left(\frac{(\sum D)^2}{N}\right)}{(N-1)(N)}}}$$

Case 3: A/B Testing: Usability Testing

- Langkah apa yang akan anda lakukan berdasarkan hasil Usability Testing?

Next Action
a. Launch tampilan aplikasi terbaru b. Lakukan A/B testing dengan variabel/metri yang lain c. Depth interview secara acak peserta usability testing demi mendapat masukan improvement

Regression Analysis

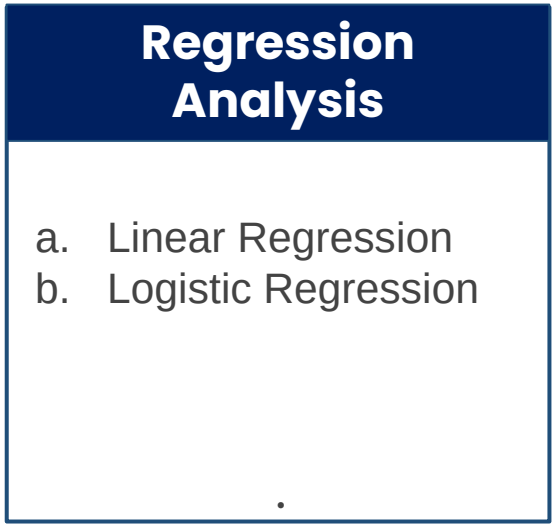
- Metode statistik yang digunakan untuk menggambarkan fungsi hubungan antara variabel dependen (y) dengan satu atau lebih variabel dependen (x)
- Tujuannya adalah untuk mengetahui hubungan sebab-akibat antara variabel x dengan y, serta melakukan prediksi nilai y dengan menggunakan persamaan regresi.
- Salah satu penerapan regression analysis adalah marketing cost analysis, yaitu mengukur hubungan antara biaya marketing yang dikeluarkan dengan output yang dicapai, serta memprediksi output dari biaya marketing di masa mendatang sebagai bahan pertimbangan dalam menyusun strategi marketing kedepannya.

Case 4: Regression - Marketing Cost Analysis

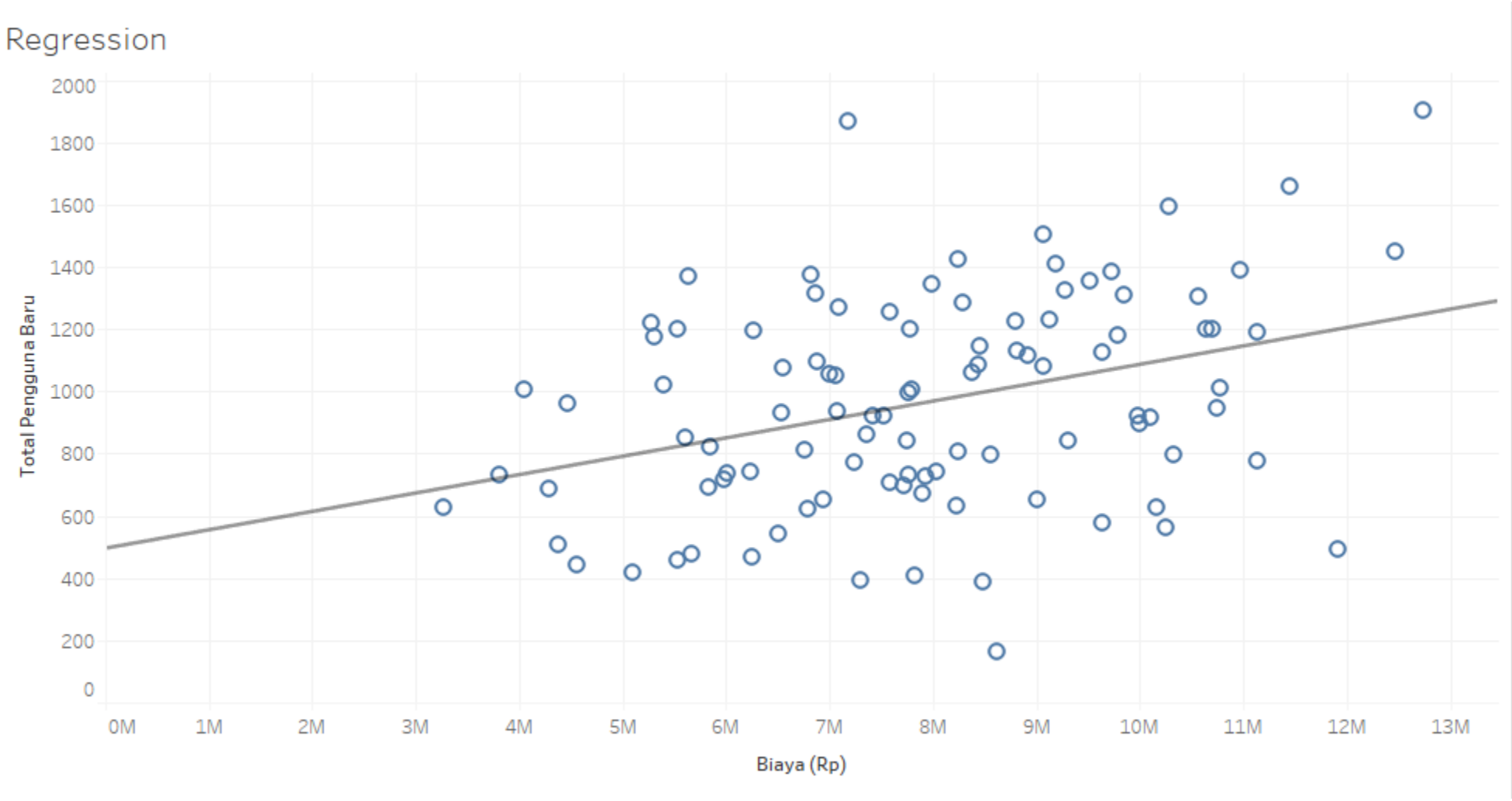
- Anda adalah data analis di tim marketing perusahaan besar di bidang oil & gas. Perusahaan tersebut berencana merilis aplikasi yang dapat digunakan untuk berbagai keperluan seperti mencari pom bensin terdekat atau membuat pengaduan terkait kelangkaan bensin.
- Karena aplikasi seperti ini masih jarang di Indonesia, perusahaan berniat melakukan marketing campaign dalam skala yang besar. Tim Anda ditugaskan untuk menjalankan marketing campaign dengan menggunakan jasa influencer.
- Setelah campaign berjalan kurang lebih seminggu, Anda mendapatkan report berupa biaya yang dikeluarkan (x) serta jumlah download aplikasi dari jasa influencer tersebut (y).

Case 4: Regression - Marketing Cost Analysis

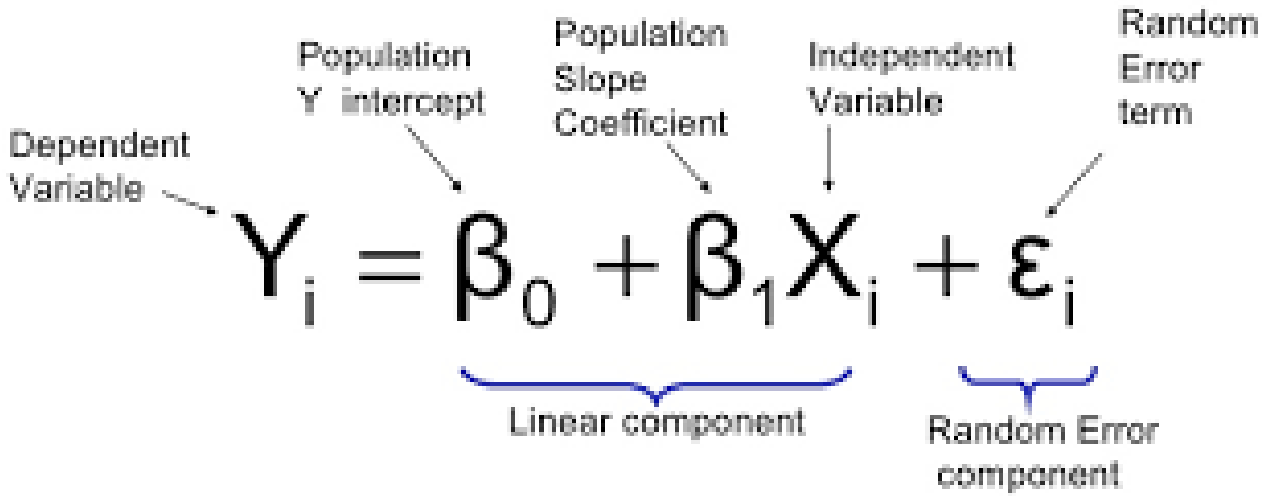
- Dengan menggunakan statistika inferensia, metode apakah yang cocok buat Anda dalam menganalisa hubungan antara biaya yang dikeluarkan (x) dan jumlah download aplikasi dari jasa influencer tersebut (y)?



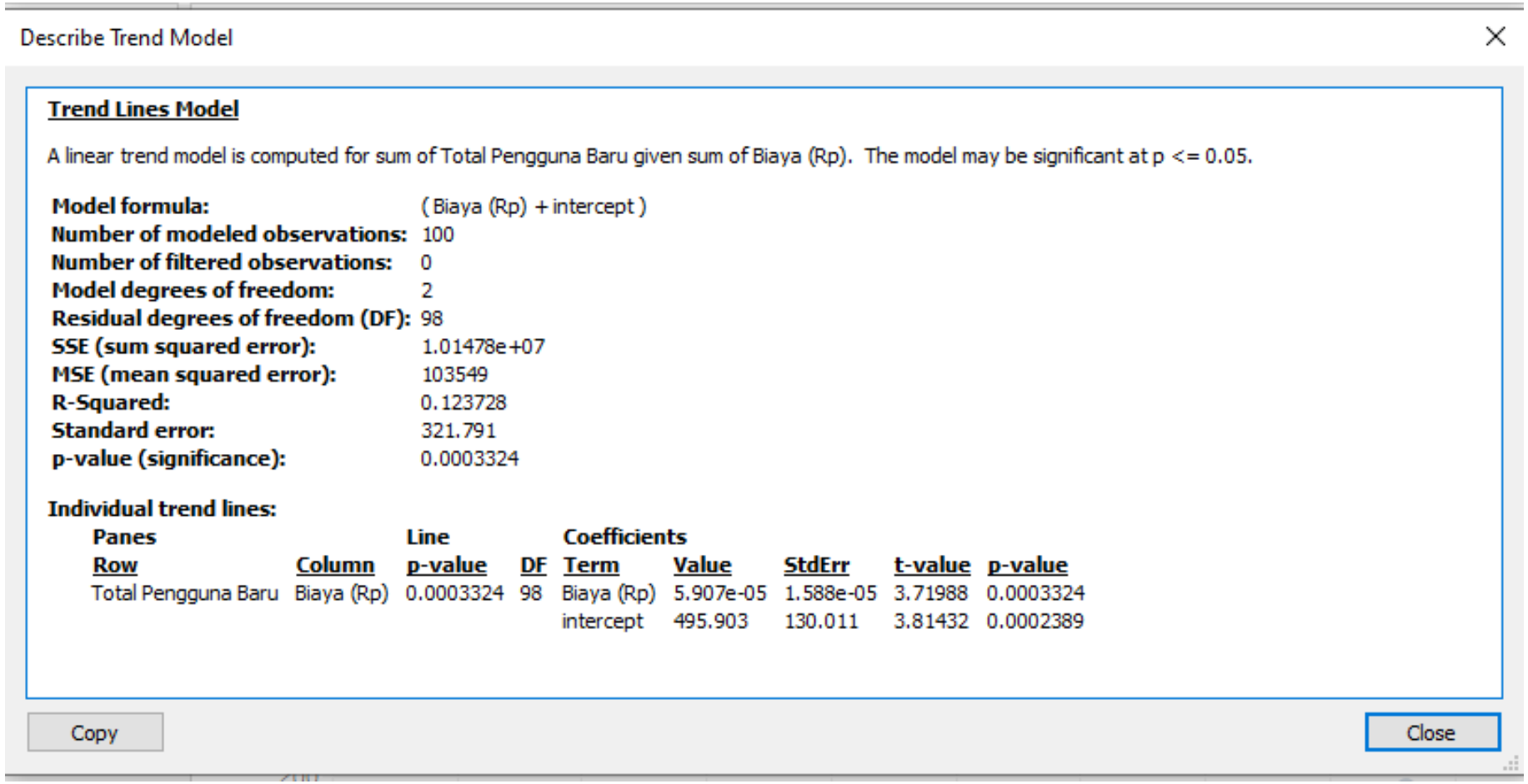
Case 4: Regression - Marketing Cost Analysis



Case 4: Regression - Marketing Cost Analysis



Case 4: Regression Analysis - Marketing Cost Analysis



Case 4: Regression - Marketing Cost Analysis

- Langkah apa yang akan anda lakukan berdasarkan hasil Marketing Cost Analysis?

Next Action
a. Melanjutkan program campaign dengan influencer existing (dimana harga dan outputnya berkorelasi positif dan secara uji statistic signifikan) b. Melakukan uji coba lanjutan dengan menyeleksi influencer yang ouput per value nya lebih tinggi. c. Melakukan analisis dengan variabel lain agar bisa mendapatkan persamaan regresi dengan R-square terbaik

Thank You