PROGRAM

DATA SCIENCE [DAY-5]

FGA KOMINFO 2022

RULE KELAS

FEEL FREE TO ASK

ACF

CALM AND LEARN TOGETHER

ACTIVE

TEKNIS

LOCAL

1. Install Anaconda, jika sudah klik Anancoda Prompt dibuka as administrator
2. Ketikan berikut ini dihalaman prompt dengan update pip dengan:

```
python -m pip install --upgrade pip
```

4. Install jupyter notebook:

```
python -m pip install jupyter
```

5. Buka dengan mengetikan:

```
jupyter notebook
```

DQLab

ONLINE

Jika kamu ingin akses secara online untuk pembelajaran python dapat melalui link dibawah ini ya!

BIT.LY/BELAJARBERSAMAKOHKRIS

CLASS PROGRAM [13 Sep – 20 Okt]

PROGRAM 1

1. Introduction to Data Science with Python
2. Python Fundamental for Data Science
3. Fundamental SQL using Select Statement
4. Fundamental SQL using Function and Group By

Introduction to Data Science with Python

Python Fundamental for Data Science

Fundamental SQL with SELECT Statement

Fundamental SQL Using FUNCTION and GROUP BY

PROGRAM 2

1. Fundamental SQL using Inner Join & Union
2. Exploratory data analysis with Python for Beginner

Fundamental SQL Using INNER JOIN and UNION

Exploratory Data Analysis with Python for Beginner

PROGRAM 3

1. Fundamental Data Visualization with Python
2. Data visualization with Python Matplotlib for Beginner

Fundamental Data Visualization with Python

Data Visualization with Python Matplotlib for Beginner - Part 1

Data Visualization with Python Matplotlib for Beginner - Part 2

PROJECT CAPSTONE

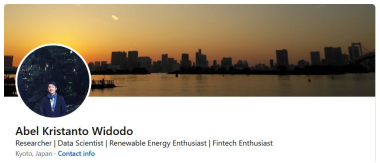
[PRESENTATION 20 Okt]

TASK

1. Peserta membuat visualisasi terkait “Data Covid19 Indonesia”
2. Peserta menyampaikan materi dalam .pptx dan .ipynb
3. Peserta dibebaskan dalam melakukan visualisasi
4. Peserta diharapkan mempelajari berikut ini:

SUBMISSION

Tuliskan ceritamu tentang hasil visualisasi yang kamu buat dan tag DQlab, Mentor di LINKEDIN



CLASS PROGRAM 2

OUR AGENDA

Exploratory data analysis with Python for Beginner

EXPLORATORY VS EXPLANATORY

DQLab

EXPLORATORY

Tujuan: identifikasi hal penting

Fokus: Melakukan pencarian

Pendekatan: Menampilkan semua metrik yang bisa di eksplorasi

EXPLANATORY

Tujuan: menggiring persepsi audience

Fokus: Komunikasi kepada audience

Pendekatan: Menyajikan chart yang mudah dibaca

EXPLORATORY DATA

Data penjualan di berbagai daerah (dalam satuan juta Rupiah)

	Toko A	Toko B	Toko C	**Total**
Aug	87	16	8	**111**
Sep	98	20	9	**127**
Okt	105	34	9	**148**
Nov	117	44	8	**169**
Des	142	54	15	**211**

INSIGT DARI SEBUAH TABEL

Data penjualan di berbagai daerah (dalam satuan juta Rupiah)

	Toko A	Toko B	Toko C	**Total**
Aug	87	16	8	**111**
Sep	98	20	9	**127**
Okt	105	34	9	**148**
Nov	117	44	8	**169**
Des	142	54	15	**211**

INSIGT DARI SEBUAH TABEL

Data penjualan di berbagai daerah (dalam satuan juta Rupiah)

	Toko A	Toko B	Toko C	**Total**
Aug	87	16	8	**111**
Sep	98	20	9	**127**
Okt	105	34	9	**148**
Nov	117	44	8	**169**
Des	142	54	15	**211**

1. Rekor penjualan tertinggi terjadi di Bulan Desember, 211 juta

INSIGT DARI SEBUAH TABEL

Data penjualan di berbagai daerah (dalam satuan juta Rupiah)

	Toko A	Toko B	Toko C	**Total**
Aug	87	16	8	**111**
Sep	98	20	9	**127**
Okt	105	34	9	**148**
Nov	117	44	8	**169**
Des	142	54	15	**211**

1. Rekor penjualan tertinggi terjadi di Bulan Desember, 211 juta
2. Toko A memiliki penjualan tertinggi dibandingkan toko lainnya

INSIGT DARI SEBUAH TABEL

Data penjualan di berbagai daerah (dalam satuan juta Rupiah)

	Toko A	Toko B	Toko C	**Total**
Aug	87	16	8	**111**
Sep	98	20	9	**127**
Okt	105	34	9	**148**
Nov	117	44	8	**169**
Des	142	54	15	**211**

1. Rekor penjualan tertinggi terjadi di Bulan Desember, 211 juta
2. Toko A memiliki penjualan tertinggi dibandingkan toko lainnya
3. Toko C memiliki penjualan yang stagnan dari Aug-Nov dan meningkat di Des

INSIGT DARI SEBUAH TABEL

Data penjualan di berbagai daerah (dalam satuan juta Rupiah)

	Toko A	Toko B	Toko C	**Total**
Aug	87	16	8	**111**
Sep	98	20	9	**127**
Okt	105	34	9	**148**
Nov	117	44	8	**169**
Des	142	54	15	**211**

1. Rekor penjualan tertinggi terjadi di Bulan Desember, 211 juta
2. Toko A memiliki penjualan tertinggi dibandingkan toko lainnya
3. Toko C memiliki penjualan yang stagnan dari Aug-Nov dan meningkat di Des
4. Penjualan toko A setiap bulan biasanya bisa 3x lebih besar dari gabungan toko B dan toko C

VISUAL EXPLANATORY UNTUK INSIGHT 2, 3

VISUAL EXPLANATORY NO 2

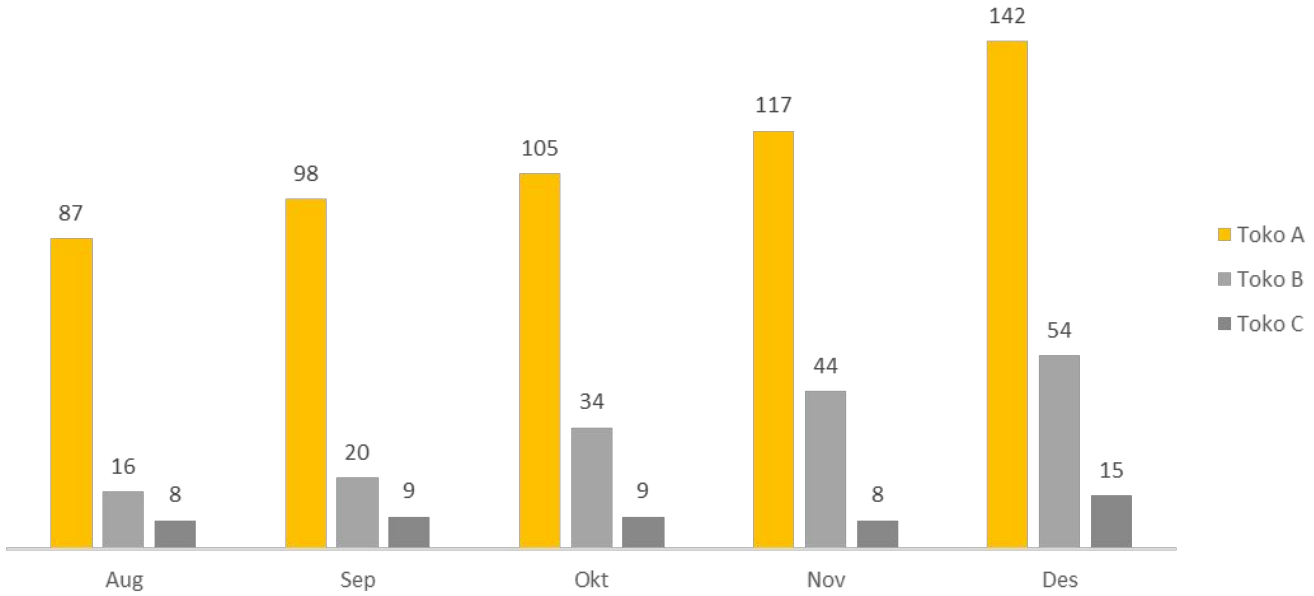
Data penjualan di berbagai daerah (dalam satuan juta Rupiah)

	Toko A	Toko B	Toko C	**Total**
Aug	87	16	8	**111**
Sep	98	20	9	**127**
Okt	105	34	9	**148**
Nov	117	44	8	**169**
Des	142	54	15	**211**

1. Rekor penjualan tertinggi terjadi di Bulan Desember, 211 juta
2. Toko A memiliki penjualan tertinggi dibandingkan toko lainnya
3. Toko C memiliki penjualan yang stagnan dari Aug-Nov dan meningkat di Des
4. Penjualan toko A setiap bulan biasanya bisa 3x lebih besar dari gabungan toko B dan toko C

OBSERVASI

Total penjualan, Aug - Dec
dalam juta rupiah



VISUAL EXPLANATORY NO 3

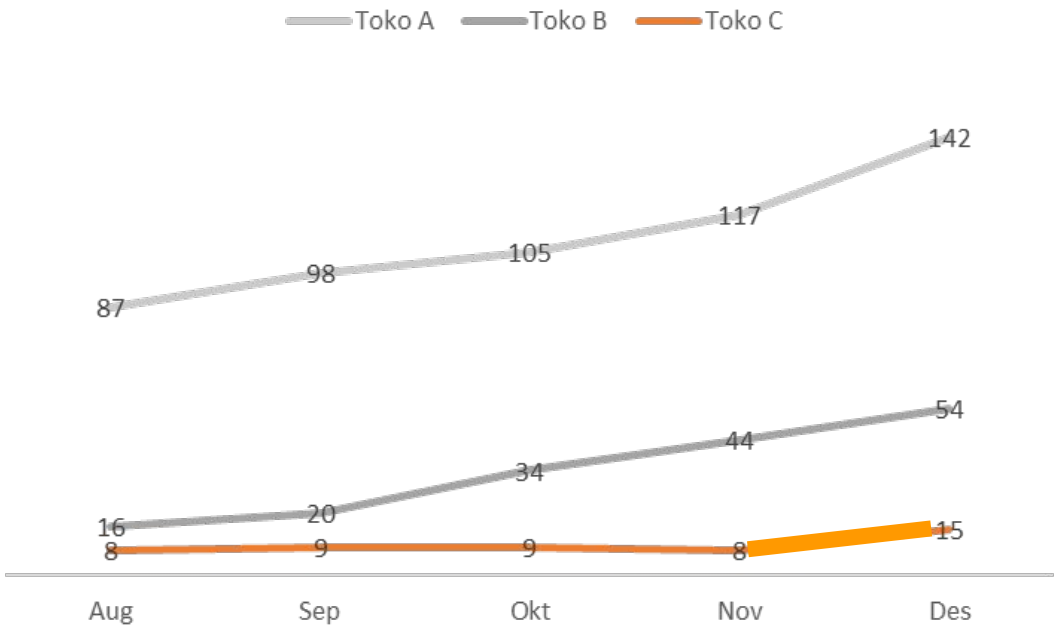
Data penjualan di berbagai daerah (dalam satuan juta Rupiah)

	Toko A	Toko B	Toko C	**Total**
Aug	87	16	8	**111**
Sep	98	20	9	**127**
Okt	105	34	9	**148**
Nov	117	44	8	**169**
Des	142	54	15	**211**

1. Rekor penjualan tertinggi terjadi di Bulan Desember, 211 juta
2. Toko A memiliki penjualan tertinggi dibandingkan toko lainnya
3. Toko C memiliki penjualan yang stagnan dari Aug-Nov dan meningkat di Des
4. Penjualan toko A setiap bulan biasanya bisa 3x lebih besar dari gabungan toko B dan toko C

OBSERVASI

Total penjualan, Aug - Dec
dalam juta rupiah



DSLab

EXPLORATORY DI PYTHON

INTRODUCTION

EDA is an approach for data analysis using variety of techniques to gain insights about the data.

Basic steps in any exploratory data analysis:

- Cleaning and preprocessing
- Statistical Analysis
- Visualization for trend analysis, anomaly detection, outlier detection (and removal).

DQLab

IMPORTANCE

Improve understanding of variables by extracting averages, mean, minimum, and maximum values, etc.

Discover errors, outliers, and missing values in the data.

Identify patterns by visualizing data in graphs such as bar graphs, scatter plots, heatmaps and histograms.

DQLab

STEPS

- Import data into workplace(Jupyter notebook, Google colab, Python IDE)
- Descriptive statistics
- Removal of nulls
- Visualization

DOLab

STEP 1 Import

Step 1 : Import pandas to the workplace.

“Import pandas”

Step 2 : Read data/dataset into Pandas dataframe. Different input formats include:

Excel : read_excel

CSV: read_csv

JSON: read_json

HTML and many more

DQLab

STEP 2.1 Stats

Used to make preliminary assessments about the population distribution of the variable.

Central tendency :

Mean – The average value of all the data points. : dataframe.mean()

Median – The middle value when all the data points are put in an ordered list: dataframe.median()

Mode – The data point which occurs the most in the dataset :dataframe.mode()

Spread : It is the measure of how far the datapoints are away from the mean or median

Variance - The variance is the mean of the squares of the individual deviations: dataframe.var()

Standard deviation - The standard deviation is the square root of the variance:dataframe.std()

Skewness: It is a measure of asymmetry: dataframe.skew()

DQLab

STEP 2.2 Stats

Used to make preliminary assessments about the population distribution of the variable.

Describe() : Summarizes the central tendency, dispersion and shape of a dataset's distribution, excluding NaN values.
Syntax: **pandas.dataframe.describe()**

Info() : Prints a concise summary of the dataframe. This method prints information about a dataframe including the index dtype and columns, non-null values and memory usage.
Syntax: **pandas.dataframe.info()**

DQLab

STEP 3 Null Values

Detecting

Detecting Null-values:

- Isnull(): It is used as an alias for dataframe.isna(). This function returns the dataframe with boolean values indicating missing values.
- Syntax : dataframe.isnull()

Handling

Handling null values:

- Dropping the rows with null values: dropna() function is used to delete rows or columns with null values.
- Replacing missing values: fillna() function can fill the missing values with a special value value like mean or median.

STEP 4 Visualization

Univariate: Looking at one variable/column at a time

- Bar-graph
- Histograms
- Boxplot

Multivariate : Looking at relationship between two or more variables

- Scatter plots
- Pie plots
- Heatmaps(seaborn)

Outlier Detection

An outlier is a point or set of data points that **lie away from the rest of the data values** of the dataset..

Outliers are easily identified by **visualizing the data**.

For example:

1. In a **boxplot**, the data points which lie outside the upper and lower bound can be considered as outliers
2. In a **scatterplot**, the data points which lie outside the groups of datapoints can be considered as outliers

Outlier Removal

Calculate the IQR as follows:

1. Calculate the first and third quartile (Q1 and Q3)
2. Calculate the interquartile range, **IQR = Q3-Q1**
3. Find the lower bound which is **Q1*1.5**
4. Find the upper bound which is **Q3*1.5**
5. Replace the data points which lie outside this range.
6. They can be replaced by mean or median.

#LEARNING BY DOING

Ada Pertanyaan?