**LAPORAN**

**FINAL PROJECT DATA SCIENCE CLUB**



Disusun Oleh :

Nama : Abdullah Jabbar Daimuddin

Nim : 192400018

PROGRAM STUDI STATISTIKA

FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS PGRI ADI BUANA SURABAYA

2021/2022

**DAFTAR ISI**

Table of Contents

[1.1 DESKRIPSI DATASET 3](#_Toc94363907)

[2.1 DATA PREPOCESSING 3](#_Toc94363908)

[3.1 MODELLING dan EVALUATING 5](#_Toc94363910)

[4.1 KESIMPULAN 8](#_Toc94363918)

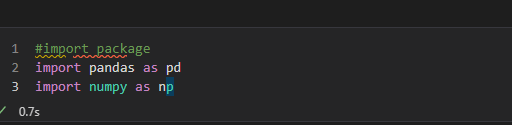
# DESKRIPSI DATASET

Dataset yang digunakan adalah data faktor yang mempengaruhi indeks pembangunan manusia di Provinsi Jawa Timur pada tahun 2020 yaitu tingkat kemiskinan, angka harapan hidup, dan pengangguran terbuka. Untuk tingkat kemiskinan (ribuan), angka harapan hidup (tahun), dan pengangguran terbuka (persen) dengan masing-masing datanya berjumlah 38.

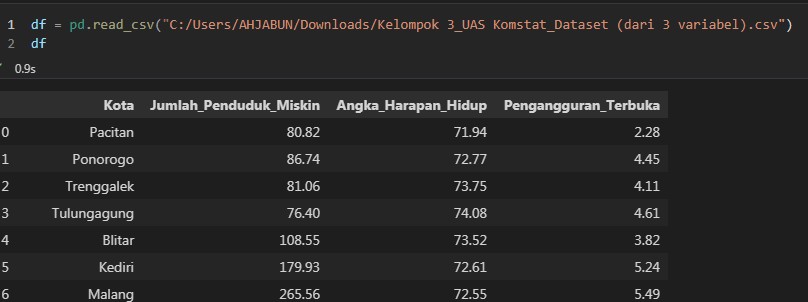
# DATA PREPOCESSING

**Load Data**

1. import package, package yang digunakan yaitu pandas dan numpy pandas untuk mengubah data csv ke dataframe

numpy untuk proses komputasi numerik

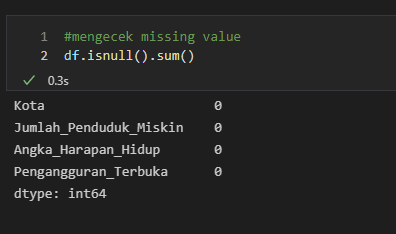
1. Load dataset yang berupa format csv menjadi format dataframe dari sebuah fungsi yaitu **pd.read\_csv** lalu diarahkan ke lokasi file yang disimpan. Lalu disimpan ke sebuah variabel yaitu **df**.

Untuk memunculkan/melihat datanya ketik variabel **df**

1. Fungsi df.head(10), untuk memunculkan data dari atas sebanyak sepuluh
2. Fungsi df.tail(10), untuk memunculkan data dari bawah sebanyak sepuluh
3. Fungsi df.shape untuk mengetahui ukuran dataset yang dipanggil
4. Fungsi df.describe menunjukkan statistika deskriptif dari data, seperti mean, median, modus, dll
5. Fungsi df.[“category”].value\_counts(), menghitung frekuensi pada kategori

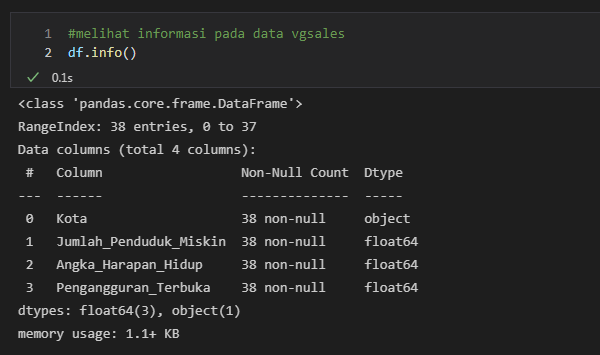
# Data Cleaning

1. Mengatasi missing value pada data, untuk mengecek missing value pada data gunakan syntax **df.isnull().sum()**, **df** yaitu variabel yang menyimpan data Kelompok 3\_UAS Komstat\_Dataset (dari 3 variabel).csv, **isnull()** untuk mengecek apakah dataframe ada nilai null dan untuk **sum()** menjumlahkan dataframe yang tidak null



Output diatas menandakan tidak ada missing value pada ketiga variabel karena nilainya nol/tidak ada missing value, maka tidak diperlukan pengisian missing value

1. Melihat informasi pada data Kelompok 3, menggunakan fungsi data.info(). Untuk datanya menggunakan variabel **df** dan fungsi dari info() yaitu untuk melihat informasi pada data



Hasil ouput diatas tidak ada Dtype yang perlu diganti. Jika variabel tersebut Dtypenya tidak cocok, maka dilakukan penggantian Dtype untuk contoh syntax **df['Kota']=df['Kota'].astype(Dtype), Dtype** yaitu Dtype baru

# MODELLING dan EVALUATING

* 1. Import Library
     1. Import numpy as np

operasi komputasi tipe data numerik seperti penjumlahan, pengurangan, perkalian, pangkat, dan operasi lainnya yang bisa diterapkan pada vektor atau matriks

* + 1. Import pandas as pd

Digunakan untuk membuat tabel, mengubah dimensi data, mengecek data, dan lain sebagainya. Struktur data dasar pada Pandas dinamakan DataFrame, yang memudahkan kita untuk membaca sebuah file dengan banyak jenis format seperti file .txt, .csv, dan .tsvd

* + 1. Import statsmodels

Untuk mengeksplorasi data, memperkirakan model statistik, dan melakukan tes statistik.

* + 1. Import patsy

untuk menggambarkan model statistik (terutama model linier, atau model yang memiliki komponen linier) dan desain matriks

* + 1. Import statsmodels.api as sm
    2. Import matplotlib.pyplot as plt Memvisualisasikan data
  1. Import Model

# from sklearn.linear\_model import LinearRegression from sklearn.cross\_validation import train\_test\_split from sklearn import metrics

* 1. Load dataset

Dataset yang digunakan yaitu ekspor.csv

**df=pd.read\_csv('Kelompok 3\_UAS Komstat\_Dataset (dari 3 variabel).csv') df**

* 1. Pilih Jumlah\_Penduduk\_Miskin dan Pengangguran\_Terbuka untuk variabel X dan Angka\_Harapan\_Hidup sebagai variabel Y **feature\_names=['Jumlah\_Penduduk\_Miskin','Pengangguran\_Terbuka'] x=df[feature\_names]**

# x y=df.Angka\_Harapan\_Hidup

* 1. Kemudian memisahkan X dan y ke dalam data latih (train) dan data pengujian (test):

# x\_train, x\_test, y\_train, y\_test = train\_test\_split(x,y,random\_state=1)

mengimpor library sklearn untuk memisahkan menjadi test set dan train set, kemudian dapat menggunakan fungsi “train\_test\_split” untuk mendefinisikan data X dan Y nya.

Misalnya data X adalah semua kolom kecuali kolom paling ujung kanan dan data Y adalah kolom paling ujung kanan.

x\_train : untuk menampung data X yang akan ditraining X\_test : untuk menampung data X yang akan ditesting Y\_train: untuk menampung data Y yang akan ditraining Y\_test : untuk menampung data Y yang akan ditesting

X dan y adalah nama variabel yang digunakan saat mendefinisikan data X dan Y.

Atribut “random\_state” untuk menghindari pembagian train set dan test set jika mengulang proses running hasilnya akan berubah-ubah. Angka yang didefinisaka random\_state bisa angka berapa saja yang berupa integer.

* 1. Ketik linear Regression Model

**Linreg=LinearRegression(),** class LiniearRegression berasal dari library sklearn.linear\_model. **LiniearRegression()** adalah fungsi untuk mengimplementasikan algoritma linear regresi di python

* 1. Membuat model dengan data train

# Linreg.fit(x\_train,y\_train)

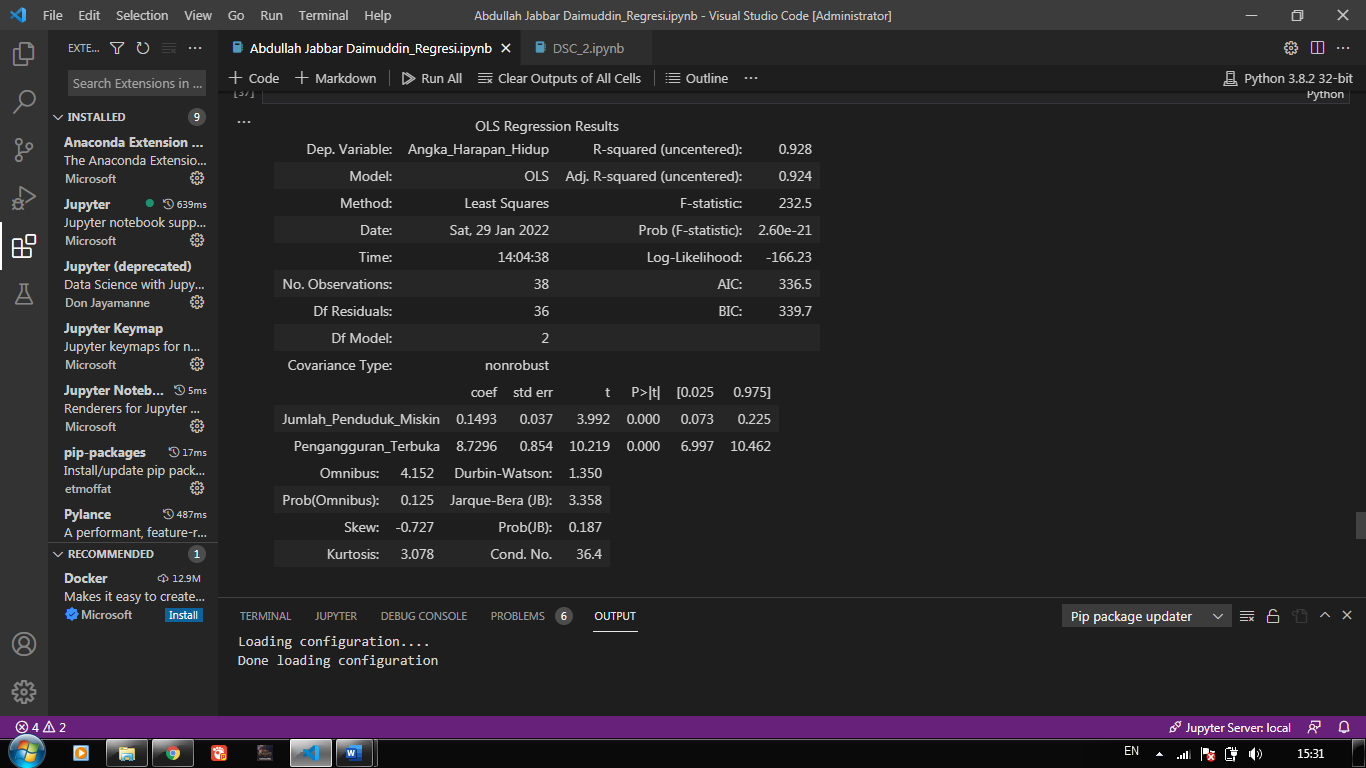
* 1. Membuat prediksi pada data pengujian

# y\_pred=Linreg.predict(x\_test)

* 1. menghitung RMSE

# print(np.sqrt(metrics.mean\_squared\_error(y\_test,y\_pred)))

* 1. mencari model regresi **model=sm.OLS(y,x).fit() predictions=model.predict(x) model.summary()**

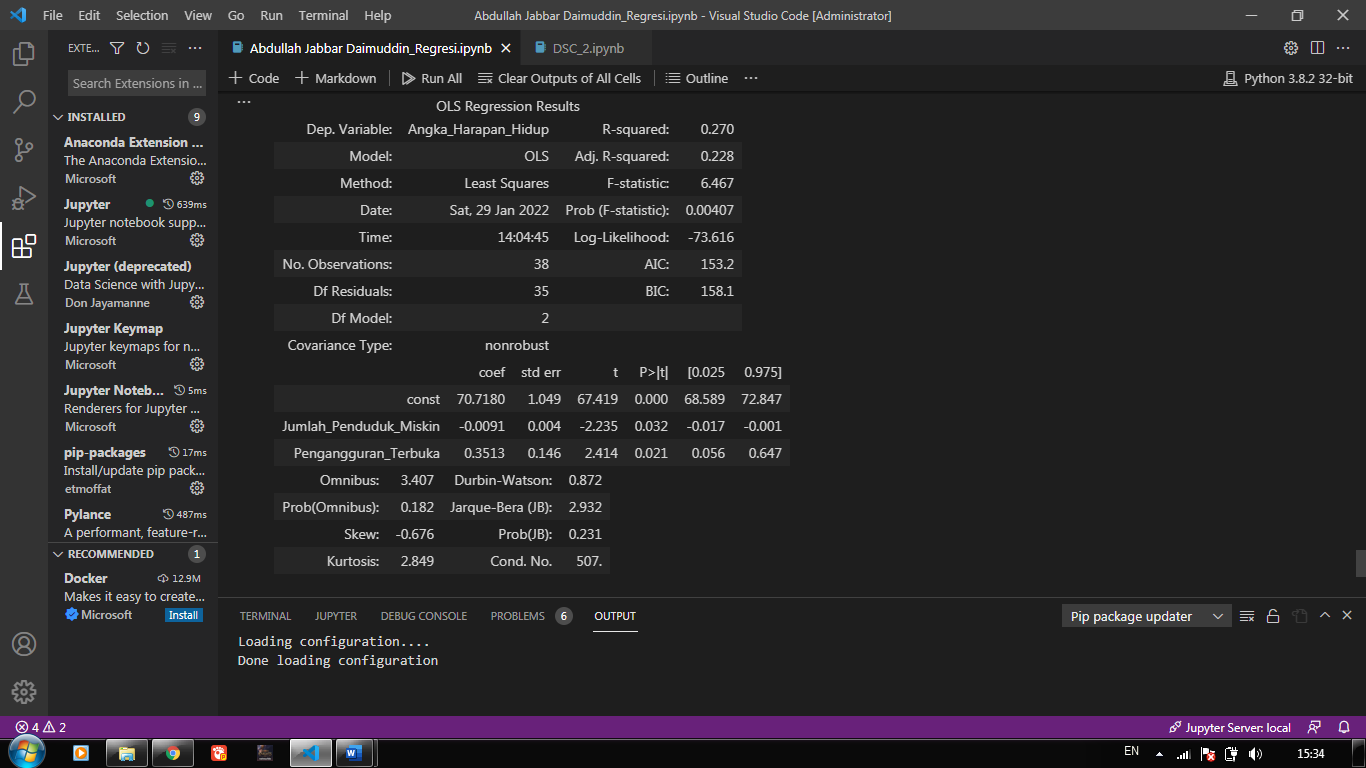


1. menambahkan variabel konstan

**x=sm.add\_constant(x)**

**model=sm.OLS(y,x).fit()**

**model.summary()**



Pada gambar  diatas diketahui bahwa fungsi yang dibuat menghasilkan model linier sederhana yaitu:

*Angka Harapan Hidup* : 70.7180 – 0.0091 *Jumlah Penduduk Miskin +* 0.3513 *Pengangguran Terbuka*

Berdasarkan model diatas artinya jika Jumlah Penduduk Miskin dan Pengangguran Terbuka mendekati nol maka nilai Angka Harapan Hidup 70.7180. Sedangkan jika Jumlah Penduduk Miskin naik satu satuan akan menurunkan nilai Angka Harapan Hidup sebesar -0.0091 dan jika Pengangguran Terbuka naik satu satuan maka akan menaikkan Angka Harapan Hidup sebesar 0.3513 dengan R-square sebesar 0.270 atau 27% variabel Jumlah Penduduk Miskin dan Pengangguran Terbuka dapat menjelaskan variabel Angka Harapan Hidup.

1. Uji Parsial

Uji parsial digunakan untuk menguji parameter secara parsial dengan kata lain untuk mengetahui apakah variable independent (X) berpengaruh secara signifikan (nyata) terhadap variable dependent (Y). Dari gambar didapat p-value (konsatanta) sebesar 0.000, nilai Jumlah Penduduk Miskin sebesar 0.032 dan nilai Pengangguran terbuka sebesar 0.021.

Hipotesis

        H0 : βi= 0, i = 0,1,2 (Tidak terdapat pengaruh secara signifikan antara X dengan Y)

      H1 : βi≠ 0, i = 0,1,2 (Ada pengaruh secara signifikan antara X dengan Y)

Tingkat signifikasi

          ∝=5% = 0.05

        Daerah kritis

        Jika p-value ≤ ∝ (0.05) → Tolak H0

         Statistik uji

P-value :  = 0.001 dan  = 0.000 ; ∝= 0,05

         Keputusan

        Karena nilai p-*value* untuk β1 dan β2 <  ∝ maka tolak

            Kesimpulan

Terdapat pengaruh secara signifikan antara variable X (Jumlah Penduduk Miskin dan Pengangguran Terbuka) dengan variabel Y(Angka Harapan Hidup).

# KESIMPULAN

# Data Prepocessing adalah suatu teknik mengolah data mentah menjadi format dan informasi yang berguna dalam hal ini terdapat 9 langkah untuk memproses data di python dengan regresi linier berganda.

# Regresi linier berganda merupakan hubungan antara dua variabel secara linier yaitu variabel indenpendent dan variabel dependen