

Deep Analytics on Student's Performances in Portuguese Class



Anugrah Budi Widhianto

Aspiring Data Analyst & Scientist

Start



Telegram
[@yourdreamid.](https://t.me/yourdreamid) ↗



Mail
Nugiewidhianto@gmail.com



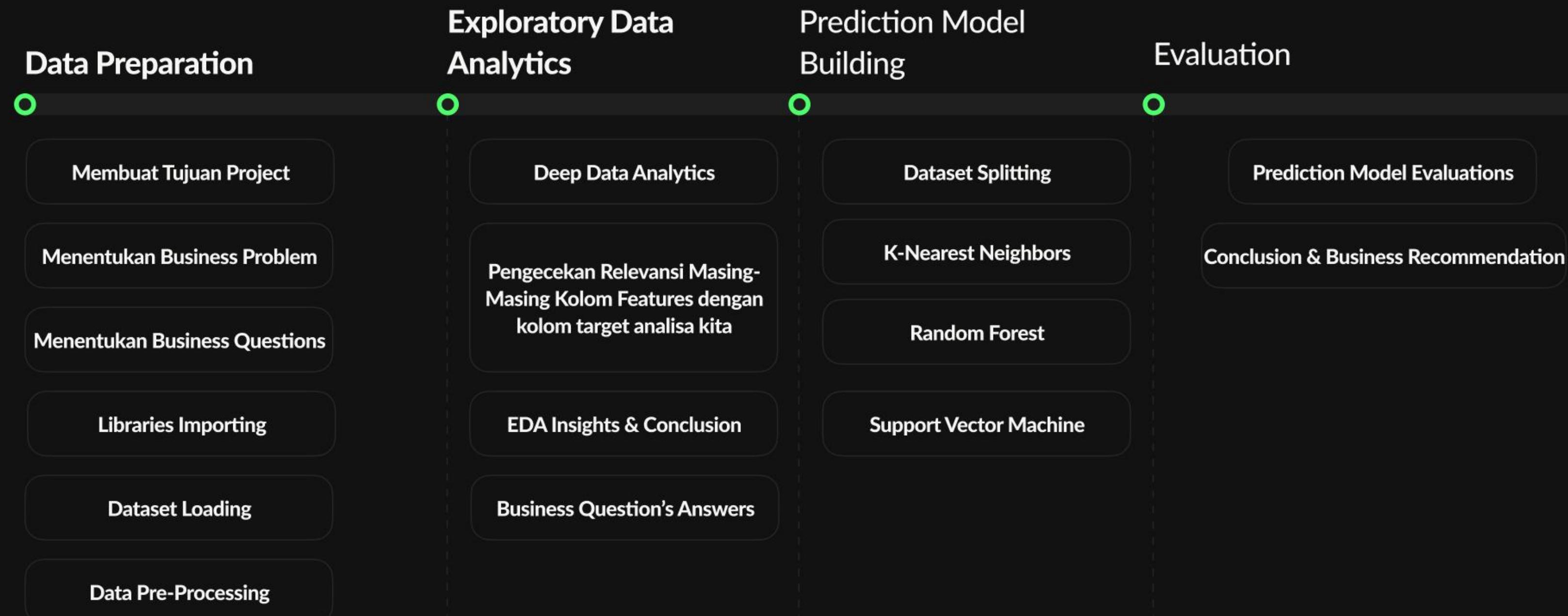
LinkedIn
[@Anugrah Budi Widhianto](https://www.linkedin.com/in/Anugrah-Budi-Widhianto) ↗

Processes

Berikut adalah proses yang saya gunakan dan implementasikan dalam pembuatan analisa data dan pembuatan model prediksi terkait final project ini

- **Finding the problem**
Melakukan komunikasi dan pengumpulan data berdasarkan kebutuhan analisa data dan pembuatan model kita.
- **Define Objectives**
Tentukan terlebih dahulu dengan jelas tujuan dan kebutuhan analisa data kita.
- **Solution Hypothesis**
Membuat hipotesis penting dalam analisis data untuk mengarahkan temuan / analisa lebih lanjut.
- **Data Pre-Processing**
Melakukan data cleaning dan processing dalam penyiapan data untuk analisa data
- **Exploratory Data Analysis**
Melakukan analisis eksploratif untuk mendapatkan wawasan awal dan memahami karakteristik data yang kita kerjakan.
- **Selecting & Applying Appropriate Technique**
Menentukan dan menggunakan teknik analisa yang tepat dan sesuai dengan tujuan dan karakteristik dari data yang kita kerjakan.
- **Interpret & Analyze Results**
Analisis hasil yang diperoleh dari temuan analisa. Interpretasikan temuan tersebut dalam konteks tujuan kita dan pertimbangkan implikasinya.
- **Communicate & Visualize Results / Findings**
Komunikasikan hasil analisis Anda dengan cara yang jelas dan mudah dipahami.
- **Evaluate, Iterate & Implement Decisions**
Evaluasi efektivitas analisis data Anda dan wawasan yang diperoleh dari analisa tersebut. Manfaatkan wawasan yang didapatkan dari analisis data untuk mengambil keputusan berdasarkan data dan mengambil tindakan yang sesuai.

Process Details & Requirements



Project Overview / Introduction

Project ini dibuat untuk mengetahui apa-apa saja faktor yang bisa menjadi pengaruh / penyebab terhadap performa siswa di sekolah menengah / SMP dan membuat **Deep Analytics** terhadap dataset dan **Model Prediksi** terhadap performa siswa disekolah dalam pelajaran portugis.

Dalam **skala bisnis**, hal ini dapat digunakan dan difungsikan dalam skala kecil maupun besar, dengan contoh:

- **Skala Kecil** : Perbaikan sistem pendidikan dan pengajaran pada suatu sekolah
- **Skala Besar** : Perbaikan sistem pendidikan, pengajaran dan pemerhatian terhadap aspek-aspek baik secara akademik ataupun diluar akademik terhadap performa siswa disekolah



About the Dataset

Dataset ini berisi **performa siswa** dalam pendidikan menengah / SMP di **dua sekolah di Negara Portugal**. Kolom-kolom pada dataset mencakup **nilai siswa, fitur demografis, sosial, dan sekolah**, yang dikumpulkan melalui **laporan sekolah dan kuesioner**. Dataset ini menyediakan informasi mengenai kinerja siswa dalam mata pelajaran Bahasa Portugis (por).



Tools & Websites

Software dan Website yang saya gunakan untuk mendapatkan dataset maupun melakukan kegiatan analisa data dan pembuatan model prediksi maupun pembuatan presentasi untuk final project ini.

Kaggle



Google Colabs



Phyton (w / libraries)



Figma



Business Problems / Questions

1. Alasan / pengaruh terhadap nilai siswa?
2. Apakah jenis kelamin mempengaruhi nilai siswa?
3. Apakah Pekerjaan dan Tingkat Edukasi orang tua mempengaruhi nilai siswa?
4. Apakah kegiatan ekstrakurikuler mempengaruhi nilai siswa secara negatif?
5. Apakah Hubungan Romantis mempengaruhi nilai siswa?



Libraries Importing

Pada proses ini, kita akan melakukan importing libraries python yang nantinya akan kita perlukan untuk **kegiatan loading dataset, analisa data, visualisasi data, manipulasi data, pembuatan model prediksi, evaluasi model dan kegiatan analisa data lainnya**. Libraries tersebut mencakup a

- Pandas
- Numpy
- Matplotlib
- Seaborn
- Scipy
- Scikit Learn

```
[ ] import pandas as pd
import numpy as np
pd.set_option('display.max_columns', None)
import matplotlib.pyplot as plt
import seaborn as sns
import scipy.stats as stats
from sklearn.model_selection import train_test_split, cross_val_score, GridSearchCV
from sklearn.preprocessing import StandardScaler, MinMaxScaler

from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import confusion_matrix, accuracy_score, ConfusionMatrixDisplay, precision_score, recall_score, roc_curve, roc_auc_score
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import chi2
from mlxtend.feature_selection import SequentialFeatureSelector
from sklearn.metrics import classification_report
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import GridSearchCV
from sklearn.metrics import accuracy_score, confusion_matrix, ConfusionMatrixDisplay
from sklearn.svm import SVC

import warnings
warnings.filterwarnings('ignore')

[ ] from google.colab import drive
drive.mount('/content/drive')
```

Dataset Loading & Basic Data Exploration

Dataset Loading

Seperti yang sudah kita ketahui sebelumnya, dataset yang akan kita gunakan adalah dataset bernama '[portuguese.csv](#)'. Sesuai dengan informasi sebelumnya, dataset ini berisikan detail tentang performa, informasi akademik maupun non akademik siswa pada kelas bahasa portugis yang didapatkan melalui [survey](#). Pada bagian proses ini, selain melakukan dataset loading, kita juga akan menggunakan code 'df.head' untuk mengetahui bagian rows teratas pada dataset yang kita miliki. Hasilnya seperti dibawah ini :

```
[ ] # Dataset loading menggunakan pandas
df = pd.read_excel("/content/drive/MyDrive/Colab Notebooks/Portuguese.csv")
# Menampilkan 5 Value teratas pada dataset kita
df.head()
```

school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian	traveltime	studytime	failures	schoolsup	famsup	paid	activities	nursery	higher	internet	romantic	famrel
GP	F	18	U	GT3	A	4	4	at_home	teacher	course	mother	2	2	0	yes	no	no	no	yes	yes	no	no	4
GP	F	17	U	GT3	T	1	1	at_home	other	course	father	1	2	0	no	yes	no	no	no	yes	yes	no	5
GP	F	15	U	LE3	T	1	1	at_home	other	other	mother	1	2	0	yes	no	no	no	yes	yes	yes	no	4
GP	F	15	U	GT3	T	4	2	health	services	home	mother	1	3	0	no	yes	no	yes	yes	yes	yes	yes	3
GP	F	16	U	GT3	T	3	3	other	other	home	father	1	2	0	no	yes	no	no	yes	yes	no	no	4

Basic Data Exploration

Untuk melakukan pendekatan awal atau pengambilan pemahaman dasar akan dataset yang akan kita proses dan gunakan dalam project, kita harus selalu memulai analisa kita dengan melakukan **Data Exploration**. Dengan metode ini, kita akan bisa mendapatkan informasi dasar terhadap dataset kita seperti **tipe data pada suatu kolom, jumlah baris dan kolom pada dataset, dll.** Adapun code-code yang biasa digunakan pada bagian proses ini seperti :

- `df.head()` (pengcekan 5 baris teratas pada dataset)
- `df.tail()` (pengecekan 5 baris terbawah pada dataset)
- `df.sample()` (pengecekan baris random pada dataset)
- `df.shape` (pengecekan jumlah baris dan kolom)
- `df.info()` (pengecekan tipe dan jumlah data pada kolom)
- `df.describe()` (informasi tentang kolom-kolom numerik)

dan lain-lain.

```
] # Pengecekan jumlah baris dan kolom pada dataset
df.shape
(651, 33)
```

```
➊ # Pengecekan informasi dasar kolom dan type data di dataset kita
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 651 entries, 0 to 650
Data columns (total 33 columns):
 #   Column      Non-Null Count  Dtype  
--- 
 0   school      651 non-null    object 
 1   sex          651 non-null    object 
 2   age          651 non-null    int64  
 3   address     651 non-null    object 
 4   famsize     651 non-null    object 
 5   Pstatus      651 non-null    object 
 6   Medu         651 non-null    int64  
 7   Fedu         651 non-null    int64  
 8   Mjob         651 non-null    object 
 9   Fjob         651 non-null    object 
 10  reason       651 non-null    object 
 11  guardian    651 non-null    object 
 12  traveltime  651 non-null    int64  
 13  studytime   651 non-null    int64  
 14  failures    651 non-null    int64  
 15  schoolsup   651 non-null    object 
 16  famsup      651 non-null    object 
 17  paid         651 non-null    object 
 18  activities   651 non-null    object 
 19  nursery     651 non-null    object 
 20  higher      651 non-null    object 
 21  internet    651 non-null    object 
 22  romantic    651 non-null    object 
 23  famrel      651 non-null    int64  
 24  freetime    651 non-null    int64  
 25  goout       651 non-null    int64  
 26  Dalc        651 non-null    int64  
 27  Walc        651 non-null    int64  
 28  health      651 non-null    int64  
 29  absences    651 non-null    int64  
 30  G1          651 non-null    int64  
 31  G2          651 non-null    int64  
 32  G3          651 non-null    int64  
dtypes: int64(16), object(17)
```

Hasil Data Exploration ‘df.describe’

# Pengecekan Numerical Value pada dataset kita df.describe()																		
	age	Medu	Fedu	traveltime	studytime	failures	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3		
count	651.000000	651.000000	651.000000	651.000000	651.000000	651.000000	651.000000	651.000000	651.000000	651.000000	651.000000	651.000000	651.000000	651.000000	651.000000	651.000000		
mean	16.745008	2.516129	2.308756	1.569892	1.930876	0.221198	3.930876	3.18126	3.187404	1.500768	2.276498	3.537634	3.654378	11.382488	11.569892	11.904762		
std	1.217609	1.134481	1.100308	0.747889	0.828241	0.592449	0.954253	1.04999	1.174824	0.923830	1.284382	1.445326	4.635853	2.777315	2.909287	3.225880		
min	15.000000	0.000000	0.000000	1.000000	1.000000	0.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000		
25%	16.000000	2.000000	1.000000	1.000000	1.000000	0.000000	4.000000	3.00000	2.000000	1.000000	1.000000	2.000000	0.000000	10.000000	10.000000	10.000000		
50%	17.000000	2.000000	2.000000	1.000000	2.000000	0.000000	4.000000	3.00000	3.000000	1.000000	2.000000	4.000000	2.000000	11.000000	11.000000	12.000000		
75%	18.000000	4.000000	3.000000	2.000000	2.000000	0.000000	5.000000	4.00000	4.000000	2.000000	3.000000	5.000000	6.000000	13.000000	13.000000	14.000000		
max	22.000000	4.000000	4.000000	4.000000	4.000000	3.000000	5.000000	5.00000	5.000000	5.000000	5.000000	5.000000	32.000000	19.000000	19.000000	19.000000		

Pada table diatas adalah contoh penggunaan code ‘df.describe’ yang diperuntukkan dalam pengecekan value-value pada numerical columns pada dataset kita. Pada table ini kita bisa mendapatkan informasi seperti jumlah baris data, rata-rata, value minimal dan juga value maximal.

Hasil Data Exploration

Value Counts & 'df.isnull'

Value Counts digunakan untuk mendapatkan jumlah masing-masing class ataupun suatu kategori pada kolom di dalam dataset kita. Pada data disamping, kita akan menggunakan value counts yang dipadukan dengan looping agar tidak perlu melakukan atau menggunakan code Value Counts pada masing-masing kolom pada dataset kita.

Kemudian, `df.isnull` digunakan untuk mengetahui / mendapatkan rows dengan value kosong pada masing-masing kolom di dataset yang kita miliki. Pada data yang saya tampilkan di sebelah kanan, saya menambahkan sedikit kode agar value null yang ditampilkan adalah presentase dari jumlah total value pada masing-masing kolom. Tapi, dapat kita lihat bahwa value null pada masing-masing kolom pada dataset kita adalah 0.

```
## Melakukan pengecekan Value Counts terhadap semua columns didalam dataset kita
for i in df.columns:
    print(i)
    print(" ")
    print(df[i].value_counts())
    print(" ")
```

```
## Melakukan pengecekan 'Null' Values pada dataset kita
df.isnull().sum().sort_values(ascending=False)/df.shape[0]*100
```

school	0.0
paid	0.0
G2	0.0
G1	0.0
absences	0.0
health	0.0
Walc	0.0
Dalc	0.0
goout	0.0
freetime	0.0
famrel	0.0
romantic	0.0
internet	0.0
higher	0.0
nursery	0.0
activities	0.0
famsup	0.0
sex	0.0
schoolsup	0.0
failures	0.0
studytime	0.0
traveltime	0.0
guardian	0.0
reason	0.0
Fjob	0.0
Mjob	0.0
Fedu	0.0
Medu	0.0
Pstatus	0.0
famsize	0.0
address	0.0
age	0.0
G3	0.0

Data Pre-Processing

Data Pre-Processing

Pada proses ini, kita akan melakukan data pre-processing sederhana.

- **Yang pertama** kita akan melakukan grouping terhadap kolom 'Grade' G1-G3 kriteria value dari 0 s/d 20 di G3 sebagai "Fail" dan dari 21 s/d 30 sebagai "Satisfied" dan dari 31 s/d 35 sebagai "Good" kemudian 36 s/d 40 sebagai "Very Good" dan yang terakhir yaitu 40 s/d 60 sebagai "Excellent".
- **Step Kedua**, yaitu saya akan melakukan grouping terhadap usia siswa.
- **Terakhir**, kita akan merubah representasi pada kolom 'failures' dan 'absences' sebagai 0 & 1.

```
[ ] df["total"] = np.array(df["G1"]) + np.array(df["G2"]) + np.array(df["G3"])

[ ] df["grades"] = pd.cut(df["total"], bins=[-1 , 20 , 30 , 35 , 40 , 60], labels=["Fail", "Satisfied" , "Good" , "Very Good" , "Excellent"])
```

Saya juga akan membuat kolom baru untuk menentukan dan menunjukkan 'Age Group'

```
[ ] df["age_group"] = np.where((df["age"] == 15), "15", np.where((df["age"] == 16), "16", np.where((df["age"] == 17), "17", "more than or equal 18")))
```

Kemudian saya akan merubah representasi pada kolom 'failures' dan 'absences' sebagai 0 & 1

```
[ ] df["failures"] = np.where(df["failures"] >0 , 1 ,0)
df["absences"] = np.where(df["absences"] != 0 , 1 , 0 )
```

Exploratory Data Analytics (EDA) & Business Question's Answers

Inisialisasi sebelum EDA

Sebelum melakukan proses EDA, kita akan menyiapkan def function bernama 'Proportions' yang berfungsi untuk melakukan group by dari kolom a dan kolom b, kemudian melakukan value counts pada masing-masing class pada 2 kolom tersebut. Def function yang akan kita buat seperti detail dibawah :

```
[ ] # Pembuatan def function Group By bernama 'Proportions' agar bisa kita gunakan berkali-kali dalam analisa kita
def proportions(column1 , column2): # Penamaan Def Functions dan parameter nya
    out = df.groupby([column1])[column2].value_counts(normalize=True) # Penggunaan function Group By
    return out.reset_index(name='count')
```

Exploratory Data Analysis (EDA) on Business Question's 1

1. Alasan / pengaruh terhadap nilai siswa?

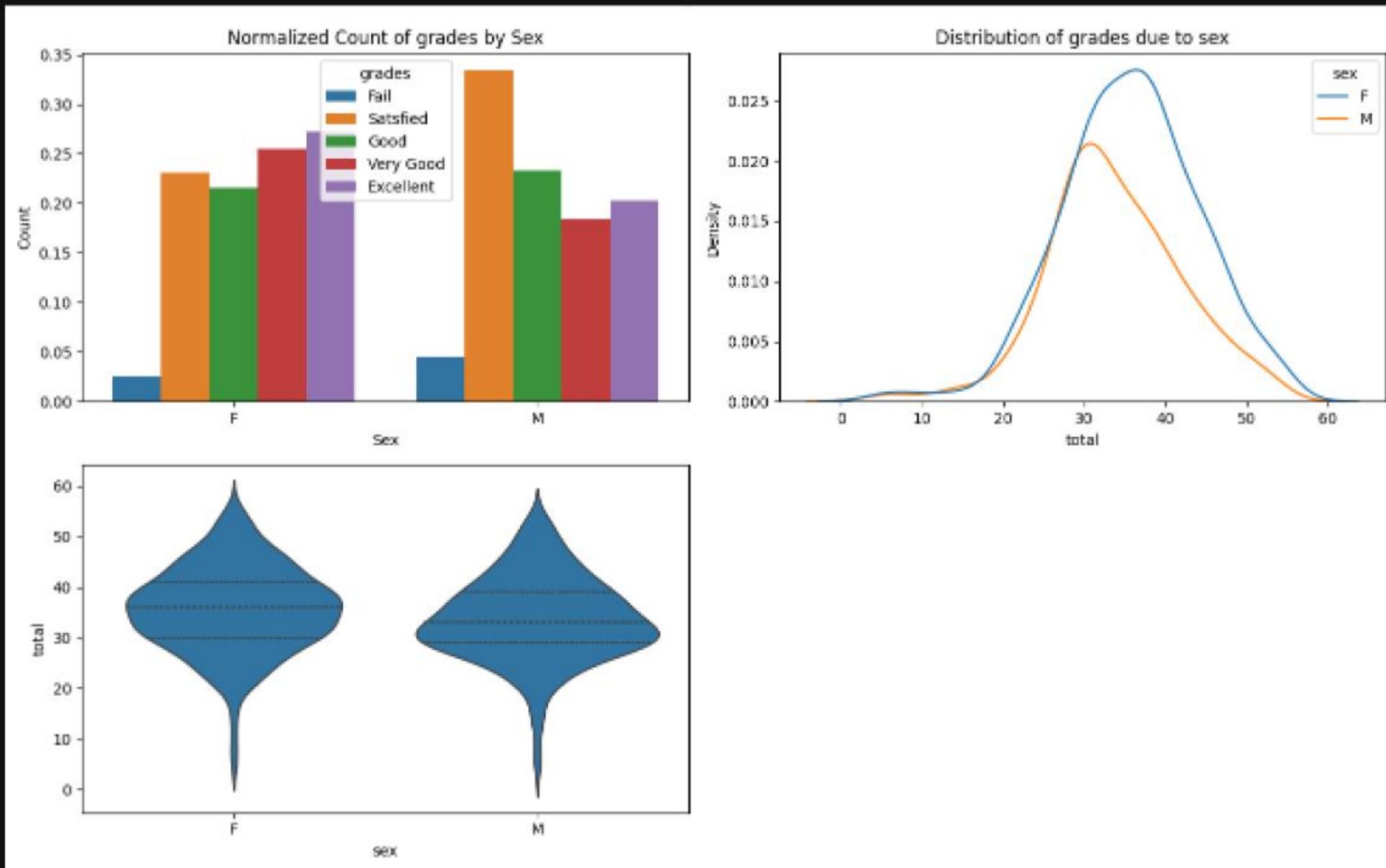
- H0 - Nilai siswa hanya dipengaruhi oleh alasan generik yang berhubungan dengan akademik
- H1 - Nilai siswa juga dipengaruhi oleh alasan diluar akademik siswa

Mengingat pertanyaan ini adalah pertanyaan yang cukup general dan mencakup banyak aspek pada dataset kita.

Pertanyaan ini akan saya lewati sebentar dan kita akan langsung loncat ke pertanyaan nomor 2 terlebih dahulu.



Exploratory Data Analysis (EDA) on Business Question's 2



2. Apakah jenis kelamin mempengaruhi nilai siswa?
- H₀ - Nilai siswa TIDAK dipengaruhi oleh jenis kelamin
 - H₁ - Terdapat PERBEDAAN NILAI antar jenis kelamin siswa

Dari hasil EDA kita pada grafik disamping, dapat kita lihat bahwa pada dataset ini ditunjukkan bahwa **siswa dengan jenis kelamin perempuan cenderung memiliki nilai-nilai yang lebih baik daripada siswa laki-laki**. Siswa laki-laki juga memiliki kecenderungan kepemilikan nilai 'fail' yang lebih tinggi

Exploratory Data Analysis (EDA) on Business Question's 3

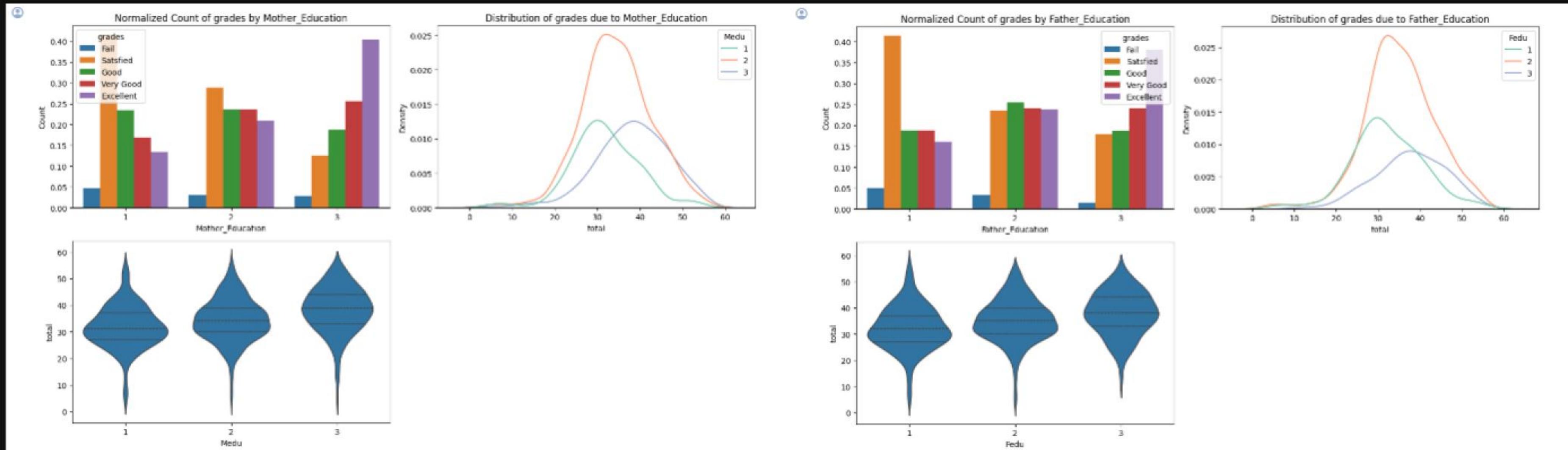
3. Apakah Pekerjaan dan Tingkat Edukasi orang tua mempengaruhi nilai siswa?

- H0 - Nilai siswa TIDAK dipengaruhi oleh Pekerjaan ataupun tingkat edukasi orang tua siswa
- H1 - Terdapat relevansi dan perbedaan yang signifikan antar nilai siswa dengan jenis pekerjaan dan tingkat pendidikan orang tua yang berbeda

Untuk Business Questions ke-3, kita memerlukan banyak grafik dan analytics karena terdapat 4 kolom yang berbeda untuk menganalisa permasalahan ini yaitu :

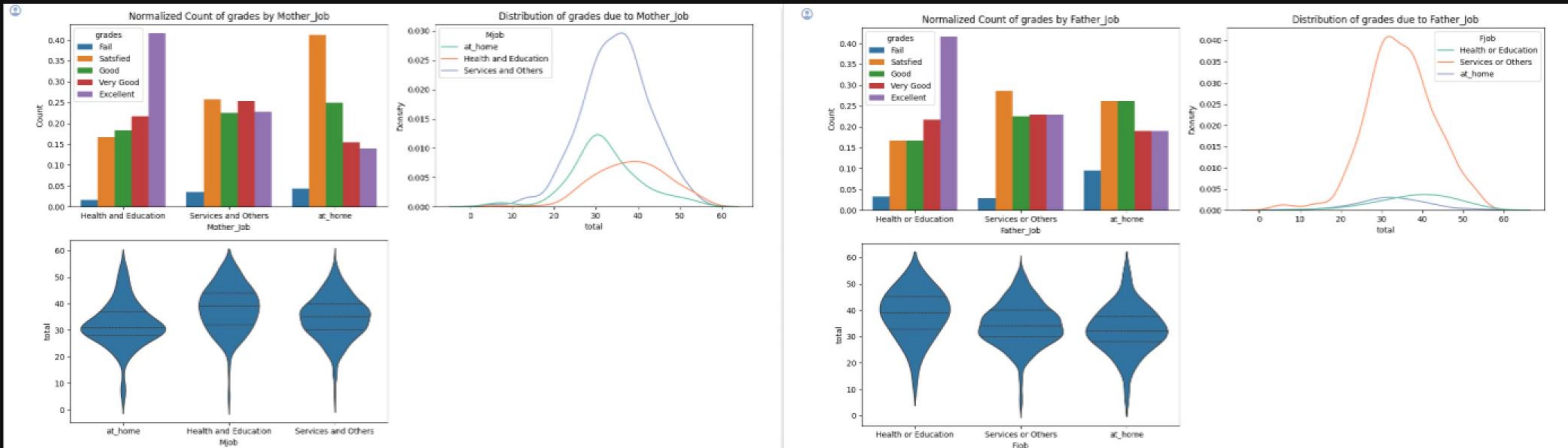
- Fedu : Father's Educations
- Medu : Mother's Educations
- Fjob : Father's Jobs / Occupations
- Mjob : Mother's Jobs / Occupations

Mother and Father's Education on Student Performance



Dapat kita lihat di visualisasi diatas, bahwa Siswa dengan tingkat 'Father & Mother's Education' yang lebih tinggi cenderung memiliki nilai yang lebih baik. Tingkat Fail dengan 'Father & Mother's Education' yang lebih rendah juga cenderung lebih besar.

Mother and Father's Jobs on Student Performance



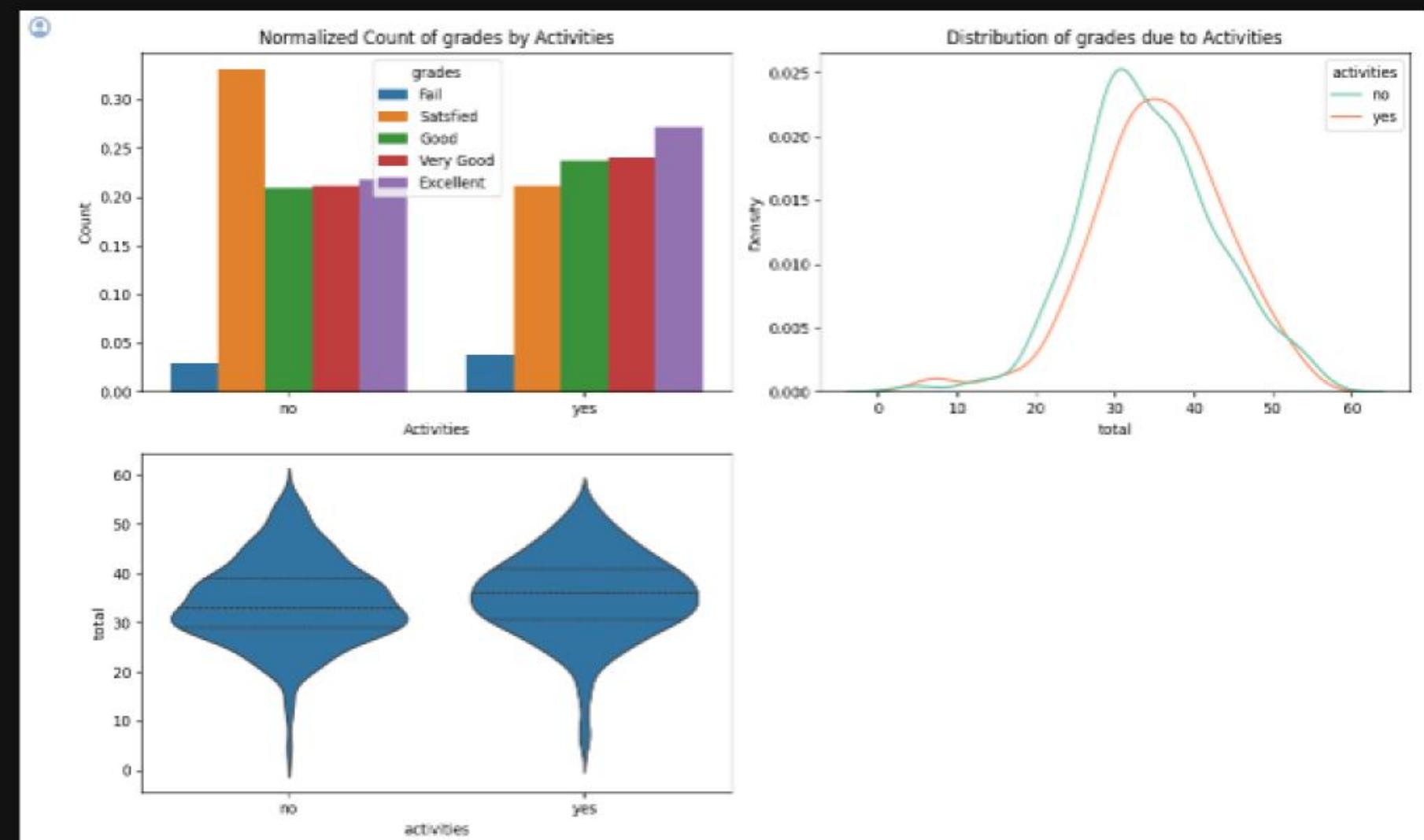
Dapat kita lihat di visualisasi diatas, bahwa Siswa yang memiliki 'Father & Mother's Jobs' dengan Class 'Health or Education' memiliki nilai yang jauh lebih tinggi dibandingkan dengan siswa lainnya. Dapat kita lihat juga bahwa class 'at-home' atau siswa dengan orangtua nya hanya tinggal dirumah cenderung memiliki nilai yang rendah dan tingkat 'fail' yang sangat tinggi

Exploratory Data Analysis (EDA) on Business Question's 4

4. Apakah kegiatan ekstrakurikuler mempengaruhi nilai siswa secara negatif?

- H0 - Kegiatan Ekstrakurikuler membuat nilai siswa menjadi lebih buruk dibanding mereka yang tidak memiliki kegiatan ekstrakurikuler
- H1 - Kegiatan Ekstrakurikuler tidak mempengaruhi nilai siswa
- H2 - Kegiatan Ekstrakurikuler membuat nilai siswa menjadi lebih baik dibanding mereka yang tidak memiliki kegiatan ekstrakurikuler

Dapat kita analisa dari grafik disamping bahwa siswa yang memiliki kegiatan ekstrakurikuler meskipun mereka memiliki tingkat fail yang sedikit lebih tinggi dari mereka yang tidak, class ini memiliki nilai yang lebih baik dan tingkat excellent yang lebih tinggi dari mereka yang tidak.

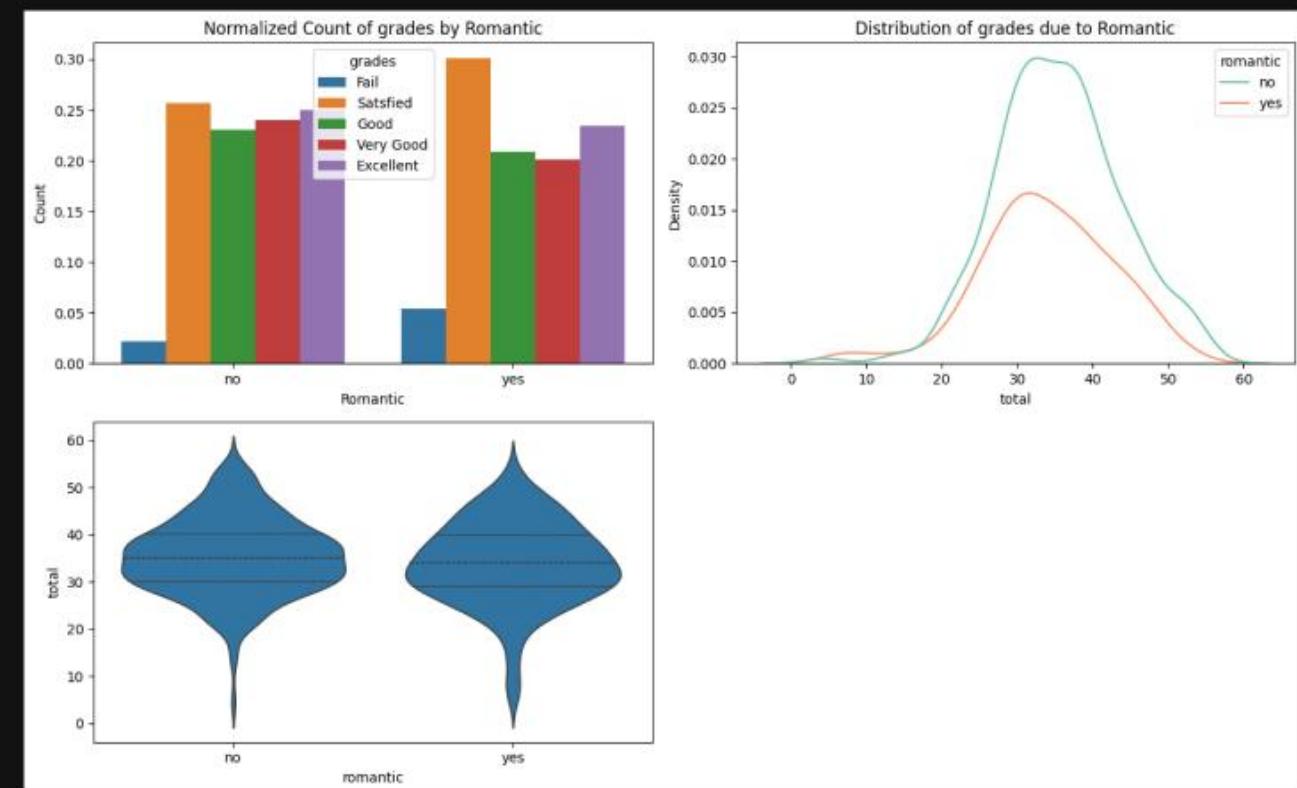


Exploratory Data Analysis (EDA) on Business Question's 5

5. Apakah Hubungan Romantis mempengaruhi nilai siswa?

- H0 - Nilai siswa TIDAK dipengaruhi oleh Hubungan Romantis siswa
- H1 - Nilai siswa DIPENGARUHI oleh Hubungan Romantis siswa

Berdasarkan Visualisasi disamping, dapat kita analisa bahwa siswa yang memiliki hubungan romantis baik didalam maupun diluar sekolah cenderung menghasilkan nilai 'fail' yang lebih banyak. Siswa-siswa yang tidak memiliki hubungan romantis juga cenderung memiliki nilai yang lebih baik dibandingkan mereka yang memiliki hubungan Romantis. Tetapi, perbandingan-perbandingan ini tidak terlalu 'signifikan', sehingga sebaiknya kita tidak mengambil kesimpulan berdasarkan fitur / faktor ini.



Exploratory Data Analysis(EDA)'s Insights & Conclusions

1. Apa sajakah **Alasan / pengaruh** terhadap nilai siswa?
2. Apakah **jenis kelamin** mempengaruhi nilai siswa?
3. Apakah **Pekerjaan dan Tingkat Edukasi orang tua** mempengaruhi nilai siswa?
4. Apakah **kegiatan ekstrakurikuler** mempengaruhi nilai siswa secara negatif?
5. Apakah **Hubungan Romantis** mempengaruhi nilai siswa?



Project's Conclusions & Recommendations

Project's Conclusions

Most Logical Hypothesis are right, hal seperti Waktu belajar; keinginan pendidikan ke jenjang lebih tinggi, dan akses internet siswa dirumah merupakan sebuah alasan kenapa siswa bisa mendapatkan nilai yang lebih baik dari siswa lain. Tetapi, tidak sedikit pula alasan lain yang ternyata memiliki pengaruh terhadap nilai siswa seperti :

- Apakah siswa mengikuti ekstrakurikuler atau tidak
- Jarak tempuh siswa dari rumah ke sekolah
- Pengasuh siswa dirumah
- Tingkat pendidikan dan pekerjaan orang tua siswa
- Alasan siswa memilih sekolah tersebut



Business Recommendations

Skala Kecil :

- Perbaikan sistem pendidikan dan pengajaran pada suatu sekolah yang tidak mendiskriminasi siswa hanya dari potensi akademik nya saja, karena terdapat banyak sekali kemungkinan hal yang bisa menyebabkan naik turun-nya nilai seorang siswa
- Mendorong pemberian dukungan terhadap siswa, baik secara akademis maupun non-akademis, karena hal-hal ini semua bisa menjadi pengaruh besar terhadap performa seorang siswa disekolah.

Skala Besar :

- Perbaikan sistem pendidikan, pengajaran dan pemerhatian terhadap aspek-aspek baik secara akademik ataupun diluar akademik terhadap performa siswa disekolah. Hal ini dapat dilakukan dan dimulai dari pembuatan sosialisasi berskala nasional di sekolah-sekolah menengah yang berfokus pada pemberian informasi tentang aspek-aspek maupun hal-hal yang bisa mempengaruhi nilai seorang siswa baik secara positif maupun negatif kepada siswa, tenaga pengajar dan juga orang tua siswa.
- Melakukan survei dengan skala dan sample yang lebih besar untuk mendapatkan hasil yang lebih signifikan dan konkret terhadap alasan-alasan yang bisa menjadi pengaruh terhadap performa siswa.





Thank You Very Much!

Please Feel Free to Ask Any Questions



Telegram
[@yourdreamid.](https://t.me/yourdreamid) ↗



Mail
Nugiewidhianto@gmail.com



LinkedIn
[@Anugrah Budi Widhianto](https://www.linkedin.com/in/Anugrah-Budi-Widhianto) ↗