LAPORAN FINAL PROJECT
BIG DATA

NAMA KELOMPOK

Albar Gusti Pamungkas	**(21.11.4122)**
Reihansyah Maulana	**(21.11.4147)**
Rayan Agil Saputra	**(21.11.4158)**
Malik Ibrahim	**(21.11.4178)**
Aldino Marsel Pratama	**(21.11.4119)**

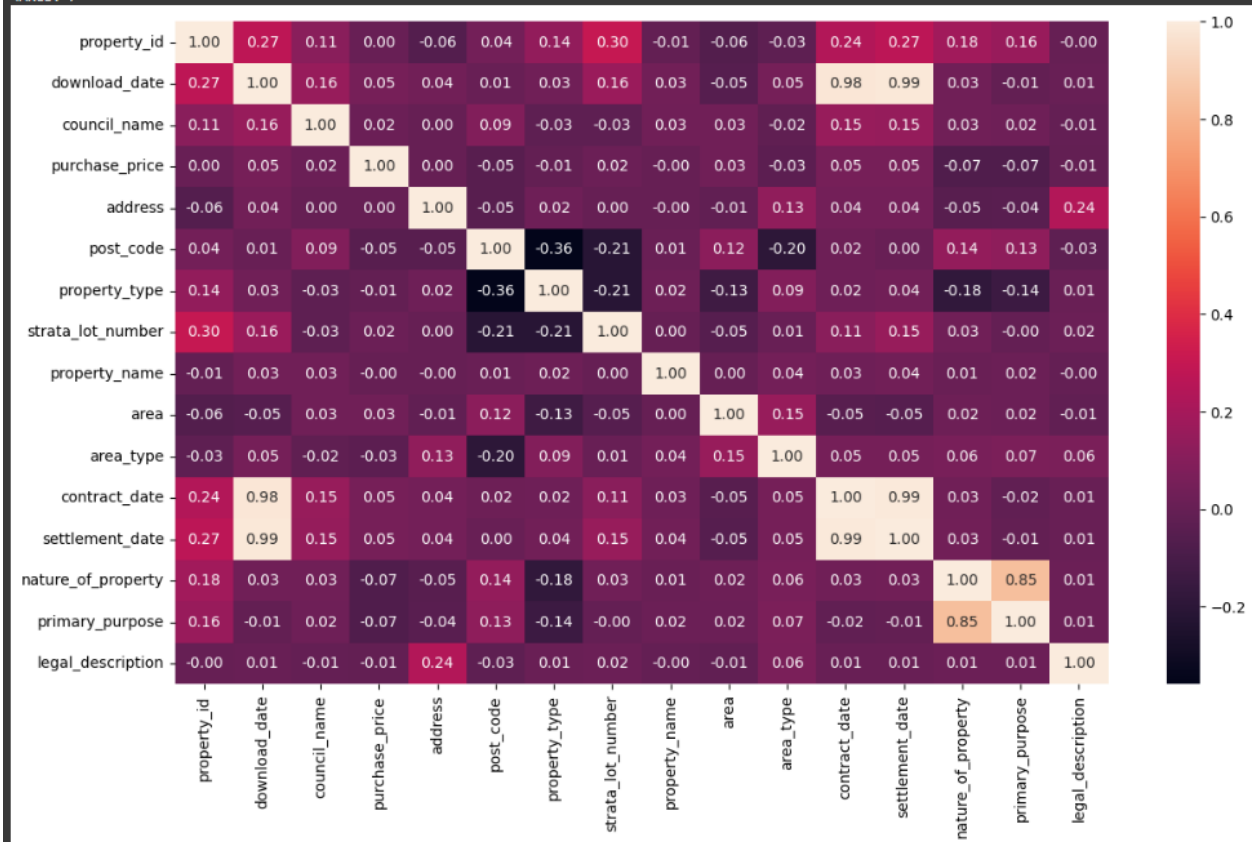
UNIVERSITAS AMIKOM YOGYAKARTA

2023

1. Heatmap

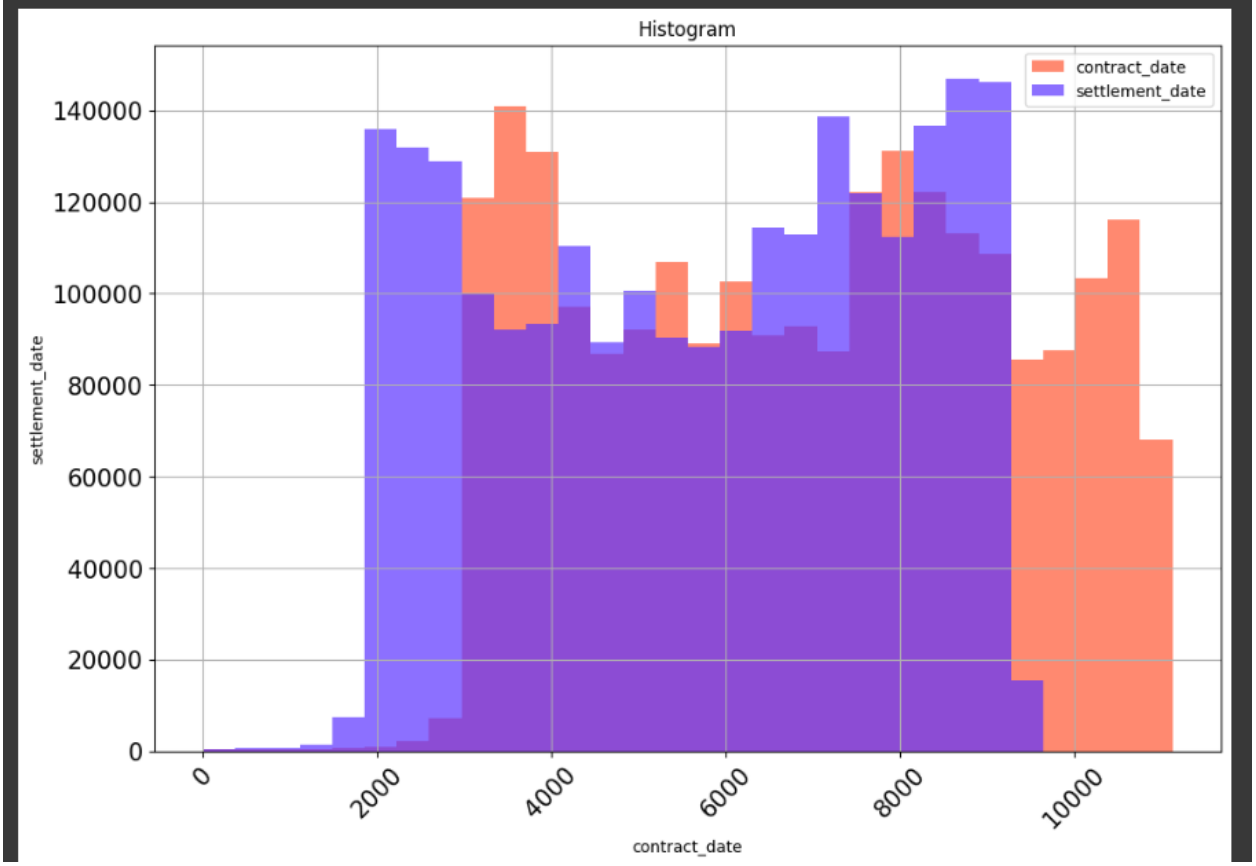
```
plt.figure(figsize=(14,8))

corr = df.corr()
sns.heatmap(corr, annot=True, fmt='.2f')
```

```
<Axes: >
```



2. Histogram

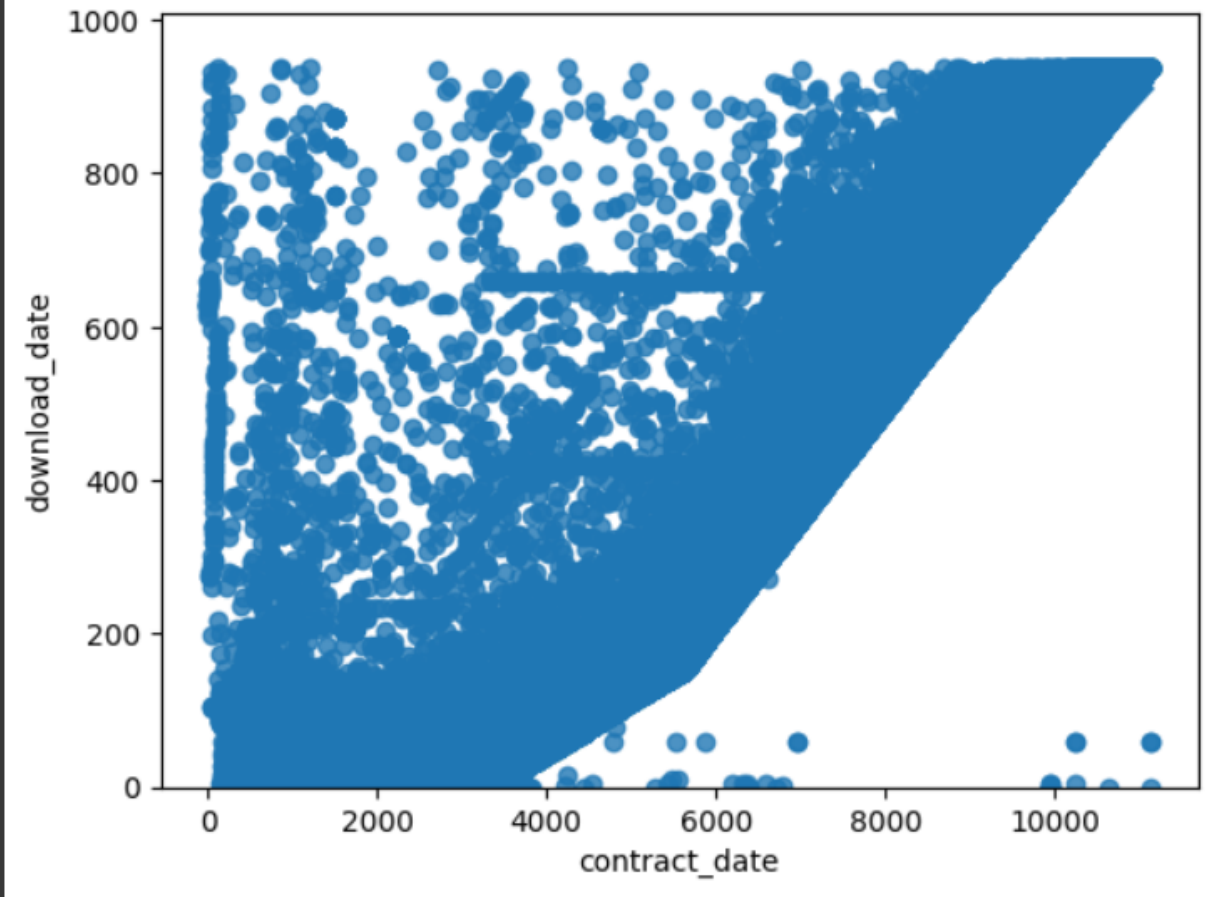
```
df[['contract_date','settlement_date']].plot(kind='hist',
alpha=0.7,
bins=30,
title='Histogram',
rot=45,
grid=True,
figsize=(12,8),
fontsize=15,
color=['#FF5733', '#5C33FF'])
plt.xlabel('contract_date')
plt.ylabel("settlement_date");
```



3. Scatter plot

```
sns.regplot(x="contract_date", y="download_date", data=df)
plt.ylim(0,)
```

```
(0.0, 1008.6640059703321)
```



4. Membuat Model Linear Regression & Evaluasi model Liniear

```
from sklearn.model_selection import train_test_split

# Membagi data dengan train test split
X_train, X_test, y_train, y_test = train_test_split(df[['contract_date', 'settlement_date']], df['download_date'], test_size=0.2, random_state=45)
```

```
train_data = pd.concat([X_train, y_train], axis=1)
test_data = pd.concat([X_test, y_test], axis=1)

# Drop rows with missing values
train_data_no_missing = train_data.dropna()
test_data_no_missing = test_data.dropna()

# Separate the target variable (y_train) and feature matrix (X_train) again
X_train_no_missing = train_data_no_missing.drop('contract_date', axis=1)  # Replace 'target_column_name' with the actual name of your target column
y_train_no_missing = train_data_no_missing['settlement_date']  # Replace 'target_column_name' with the actual name of your target column

X_test_no_missing = test_data_no_missing.drop('contract_date', axis=1)  # Replace 'target_column_name' with the actual name of your target column
y_test_no_missing = test_data_no_missing['settlement_date']  # Replace 'target_column_name' with the actual name of your target column

# Create and train the LinearRegression model with the data without missing values
lr_model = LinearRegression()
lr_model.fit(X_train_no_missing, y_train_no_missing)
```

```
▾ LinearRegression
LinearRegression()
```

```
# Menampilkan koefisien

lr_model.coef_
```

```
array([1.00000000e+00, 5.45077907e-17])
```

```
# Menampilkan intercept

lr_model.intercept_
```

```
-3.637978807091713e-12
```

```
# Menguji model

y_pred = lr_model.predict(X_test_no_missing)
y_pred
```

```
array([8193., 2829., 2480., ..., 8850., 7087., 9057.])
```

+ Code + Text

```
from sklearn.metrics import r2_score

# Menampilkan nilai r2 score
print("R2-score: %.2f" % r2_score(y_pred, y_test_no_missing))
```

```
R2-score: 1.00
```

```
from sklearn.metrics import mean_squared_error

mean_squared_error(y_test_no_missing, y_pred)
```

```
2.1266520399774333e-24
```

```
mean_squared_error(y_test_no_missing, y_pred, squared=False)
```

```
1.458304508659777e-12
```

Program

https://colab.research.google.com/drive/1ZUHNFSV79xjQpWsQ6p5axruxWOw_U2vp#scrollTo=55q3Dz_Zq6c7&line=1&uniqifier=1

Sumber Data set

https://www.kaggle.com/datasets/josephcheng123456/nsw-australia-property-data?resource=download