



Data Analytics Portfolio

Hadad Karsa

Based on Final Project Assignment at Data, Business
Analytics, & Operations Bootcamp Ruangguru

Readme First!!!


The Look merupakan E-Commerce yang menjual produk berupa pakaian. Dalam suatu perusahaan pastinya memiliki pihak internal maupun eksternal. Internal The Look terbagi menjadi 4 departement antara lain Human Resources, Sales, Marketing, dan Product. Sedangkan untuk eksternal adalah customer. **Dari keempat departemen yang ada saya bertanggungjawab terhadap analisis data departemen Human Resource/HR.**

Setiap perusahaan memerlukan insight yang tepat untuk keputusan-keputusan yang akan dibuat, begitu juga dengan perusahaan The Look. Untuk mendapatkan insight tersebut tentunya bersumber dari data yang dimiliki, bagaimana data tersebut akan diolah, dianalisis, hingga divisualisasikan.

Dalam analisis data tidak serta merta langsung melakukan analisis dari data yang ada. Perlu adanya perumusan business problem, tujuan analisis/bisnis, eksplorasi data, data cleaning, pemodelan data, hingga mendapat visualisasi yang tepat sebagai insight.



Metodologi Analisis



DEPARTEMENT Human Resource

Metodologi

DATA


1. Data employee (employees.csv) dan distribution_centers (distribution_centers.csv)
2. Berfokus pada variabel age, length_service, absent_hours, dan distribution_centers_id

TOOLS

1. SQL dengan PostgreSQL untuk melakukan join tabel dan manipulasi data..
2. Python dengan Google Colab untuk EDA, Cleaning, dan Modelling.
3. Tableau untuk visualisasi.

ANALISIS

1. Regresi Linier (Simple and Multiple)
2. Regresi Ridge
3. Regresi Polynomial



DEPARTEMENT Human Resource

IDENTIFIKASI MASALAH/PROBLEM STATEMENT

Dari semua variabel data yang ada, dapat diidentifikasi kinerja karyawan berdasarkan total jam absen dalam satu tahun dengan variabel `absent_hours`. Kinerja karyawan bisa menjadi masalah bagi perusahaan sehingga perlu dianalisis seperti apa kinerja karyawan dalam setahun terakhir dan membuat prediksi kedepannya.

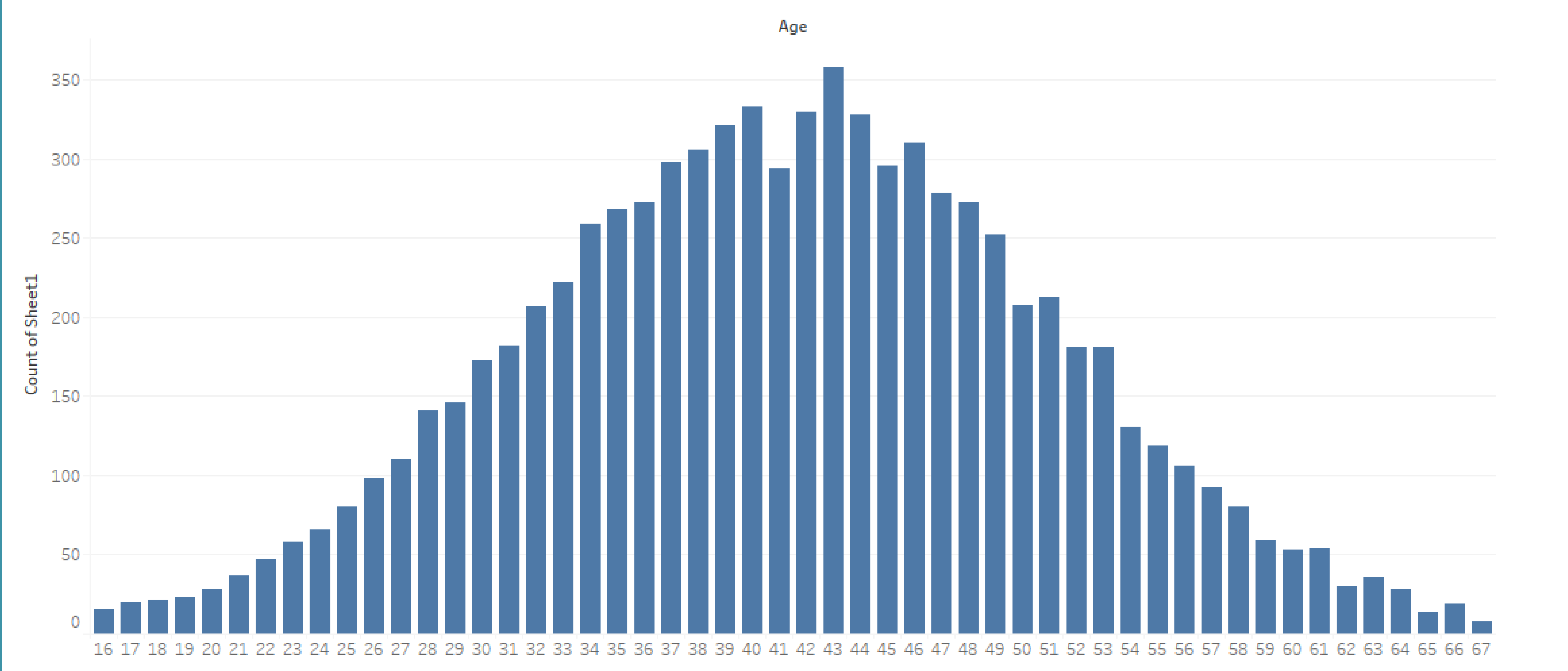
RANCANGAN SOLUSI

1. Melakukan EDA, data preparation, dan data cleaning (handling missing values dan outliers)
2. Setelah data bersih, mencari korelasi data untuk menemukan variabel yang berkorelasi dengan `absent_hours` sebagai kinerja karyawan.
3. Menentukan independent dan dependen variabel berdasarkan data correlation
4. Melakukan data analytics dan pemodelan untuk variabel independen dan dependen yang ditentukan
5. Melakukan evaluasi pemodelan dan visualisasi menggunakan variabel-variabel yang diperlukan untuk mendapatkan insight solusi.



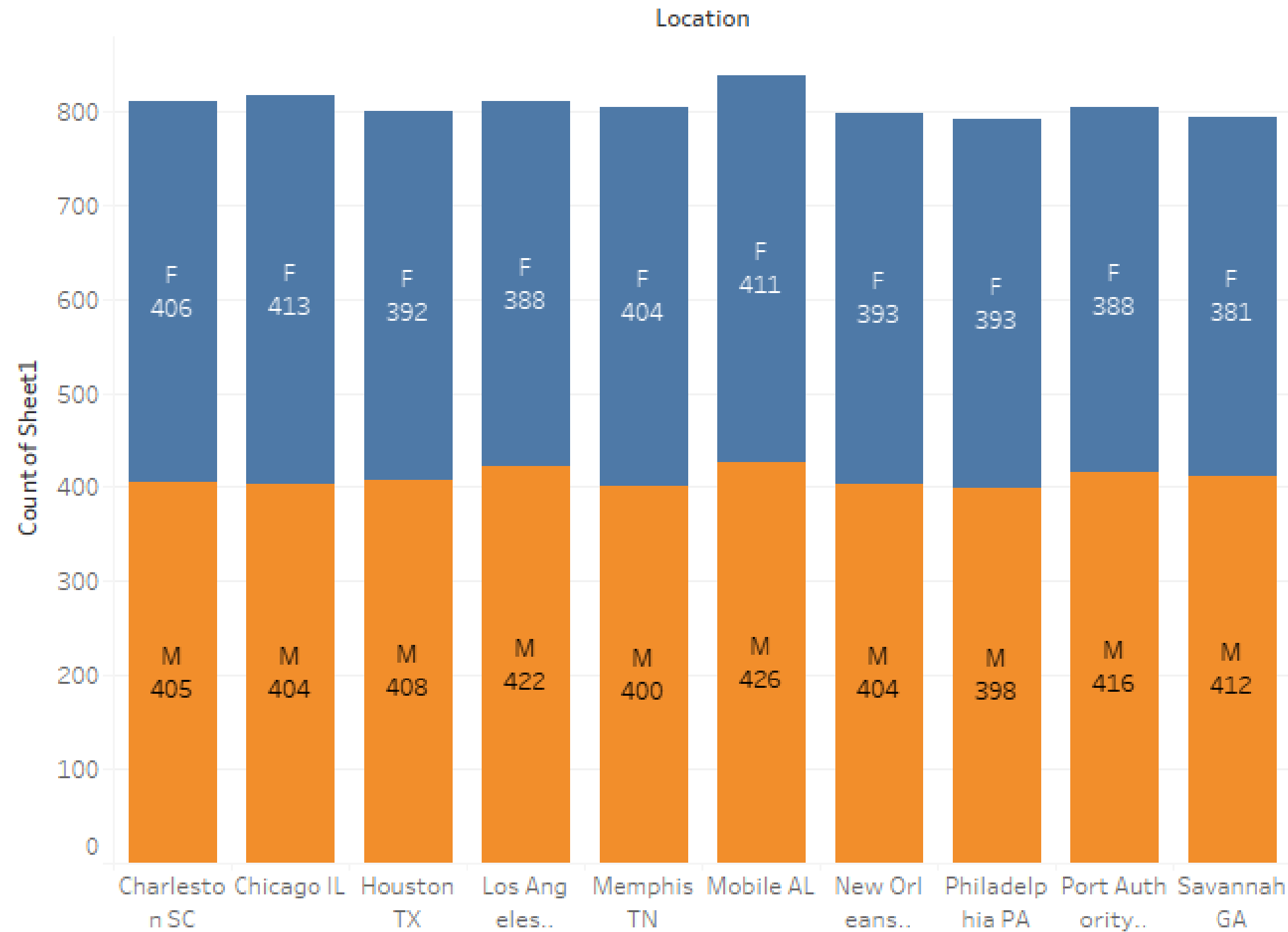
Visualisasi dengan Tableau

Distribusi Jumlah Karyawan berdasarkan Umur



Terlihat Penyebaran Umur Karyawan The Look berdasarkan jumlah karyawan terdistribusi normal

Persebaran Karyawan di Setiap Pusat Distribusi The Look



Gender

☒ (All)

☒ F

☒ M

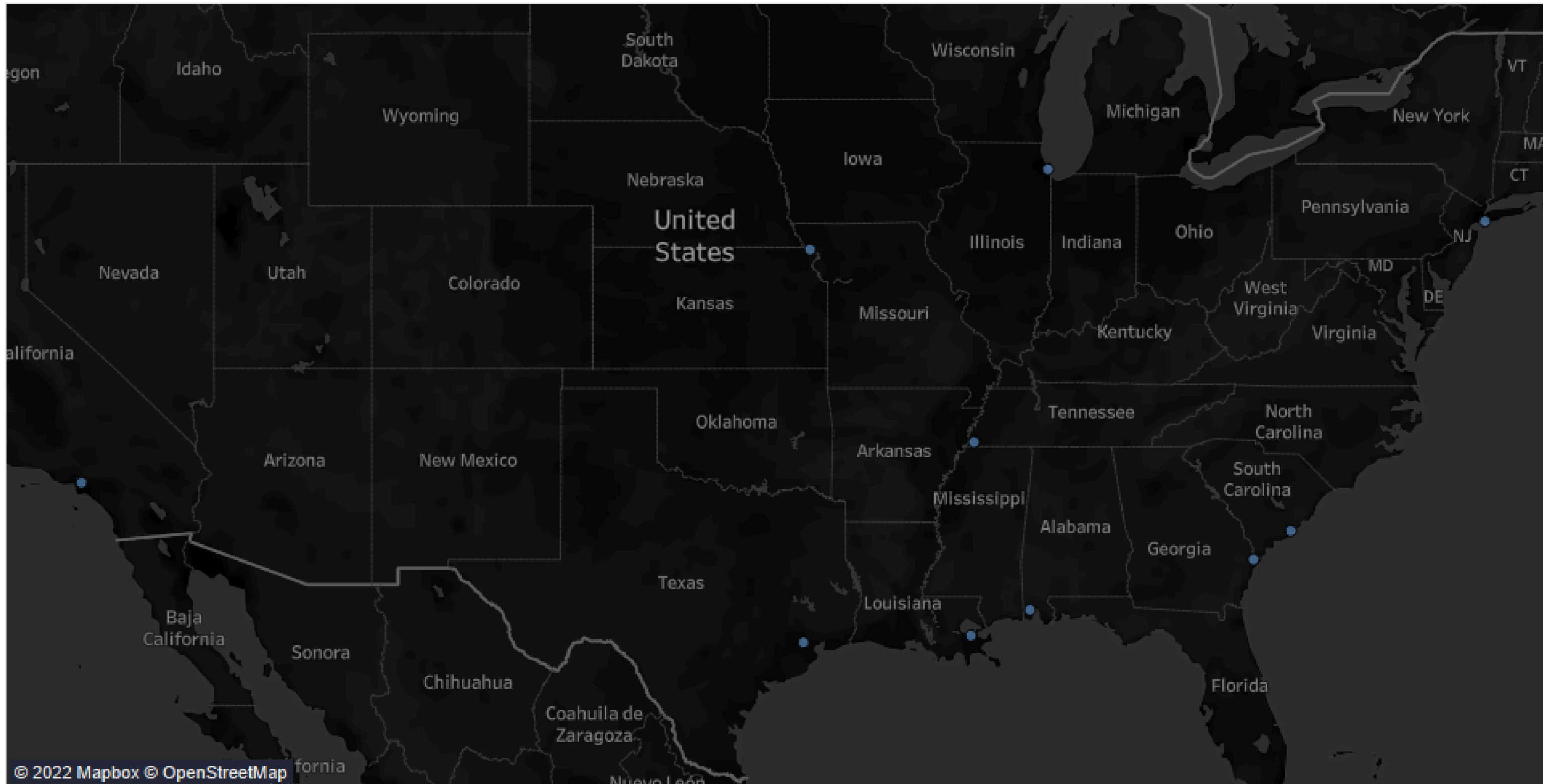
Gender

☒ F

☒ M

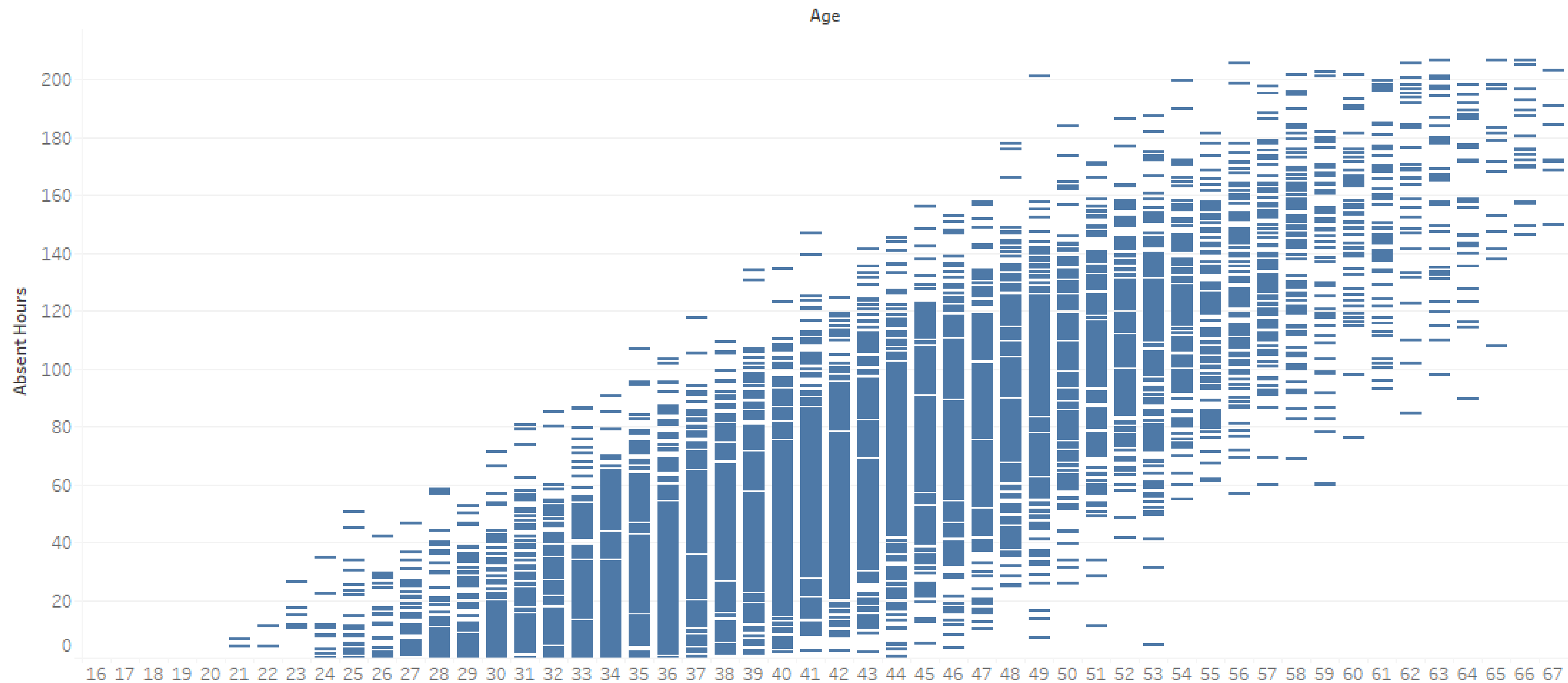
Terlihat Penyebaran jumlah karyawan di setiap lokasi kerja berdasarkan gender terdistribusi merata

Titik Lokasi Pusat Distribusi The Look (United States)



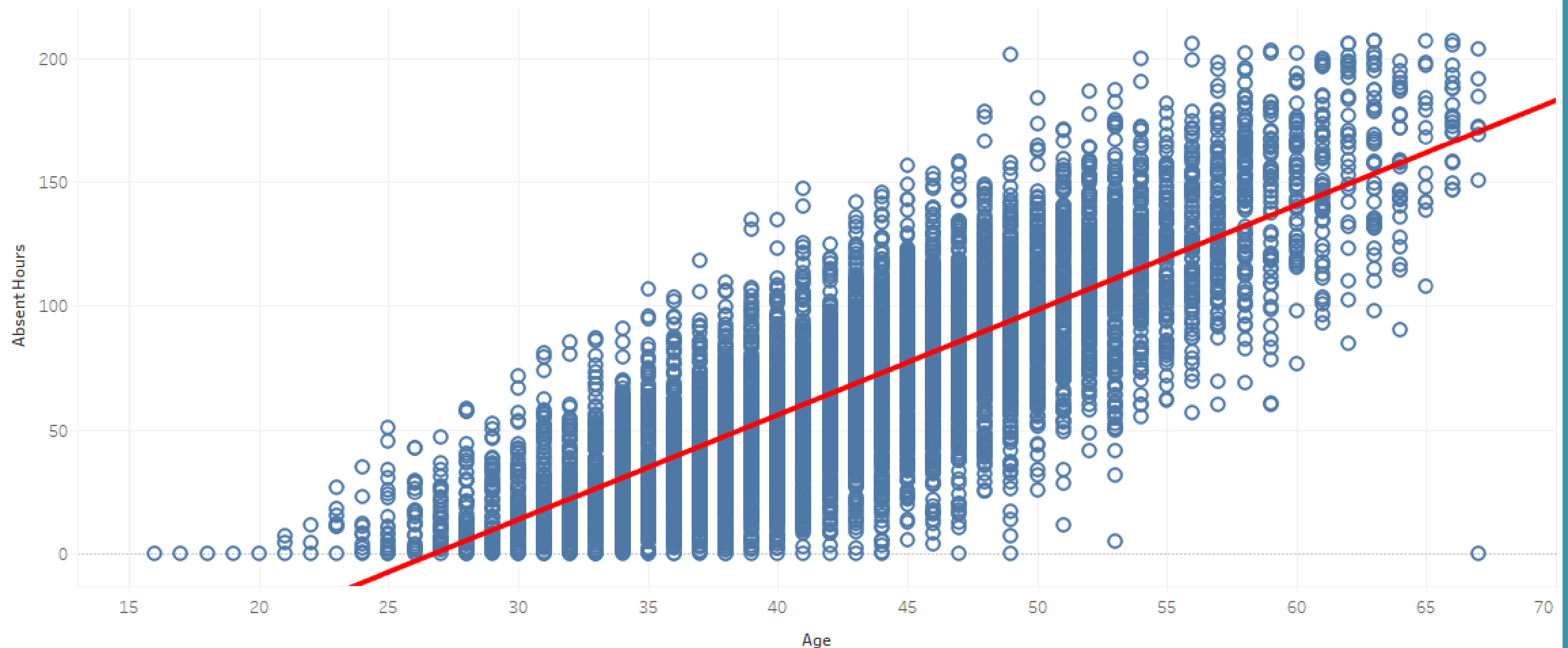
Terlihat Penyebaran titik maps lokasi pusat distribusi perusahaan The Look di Amerika

Gantt Chart Persebaran Umur Karyawan berdasarkan Jam Absen Kerja



Terlihat grafik Gantt Chart sebagai bentuk visual korelasi antara umur karyawan dan jumlah jam absen. Dimana cukup berkorelasi terhadap kinerja karyawan berdasarkan umur

Trend Lines Simple Linier Regression antara Umur dengan Jam Absen Kerja

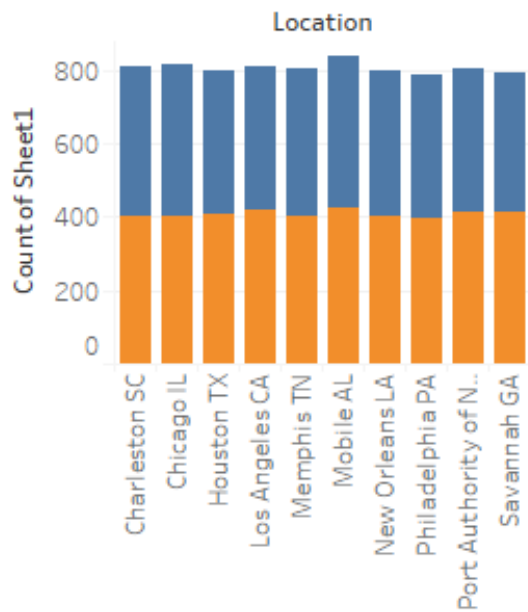


Bentuk visual lainnya yaitu Trend Lines sebagai korelasi antara umur karyawan dan jumlah jam absen. Dimana cukup berkorelasi terhadap kinerja karyawan berdasarkan umur

Dashboard

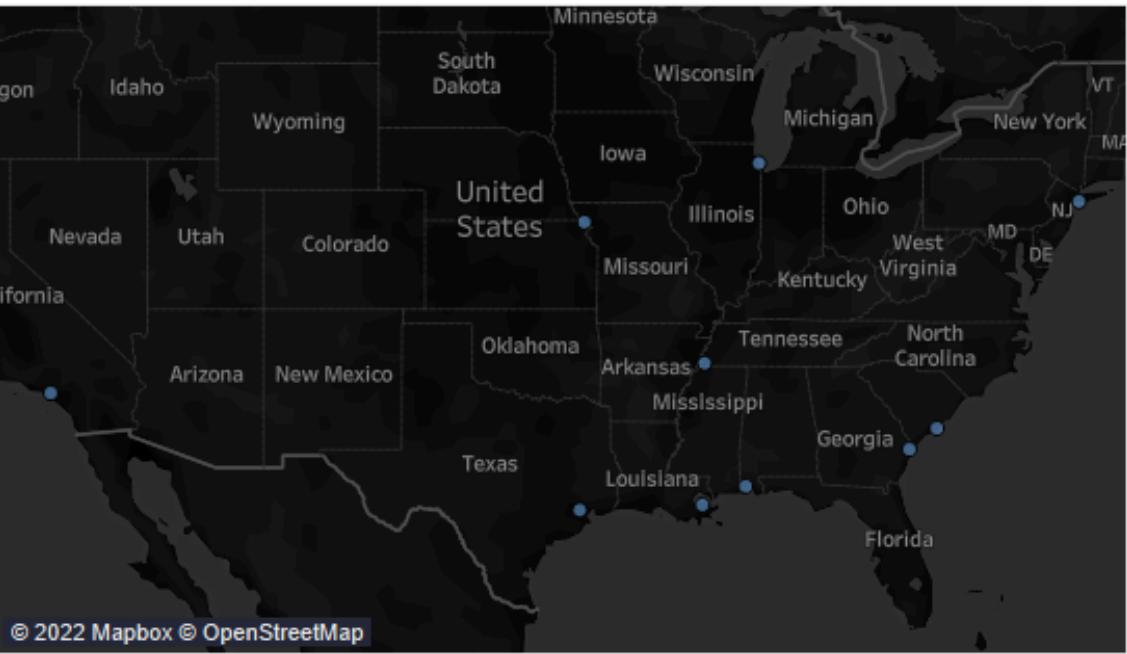
Infografis Karyawan The Look

Persebaran Karyawan di Setiap Pusat Distribusi The Look



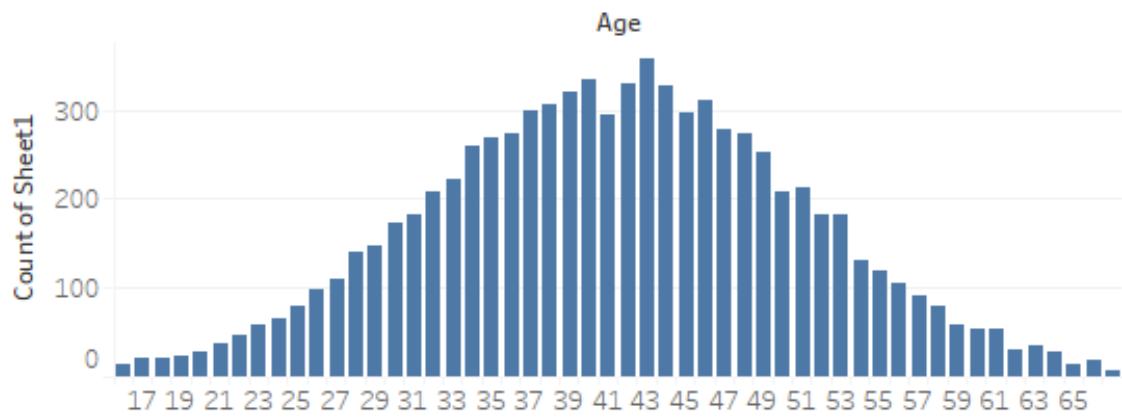
Gender
☒ F
☒ M

Titik Lokasi Pusat Distribusi The Look (United States)

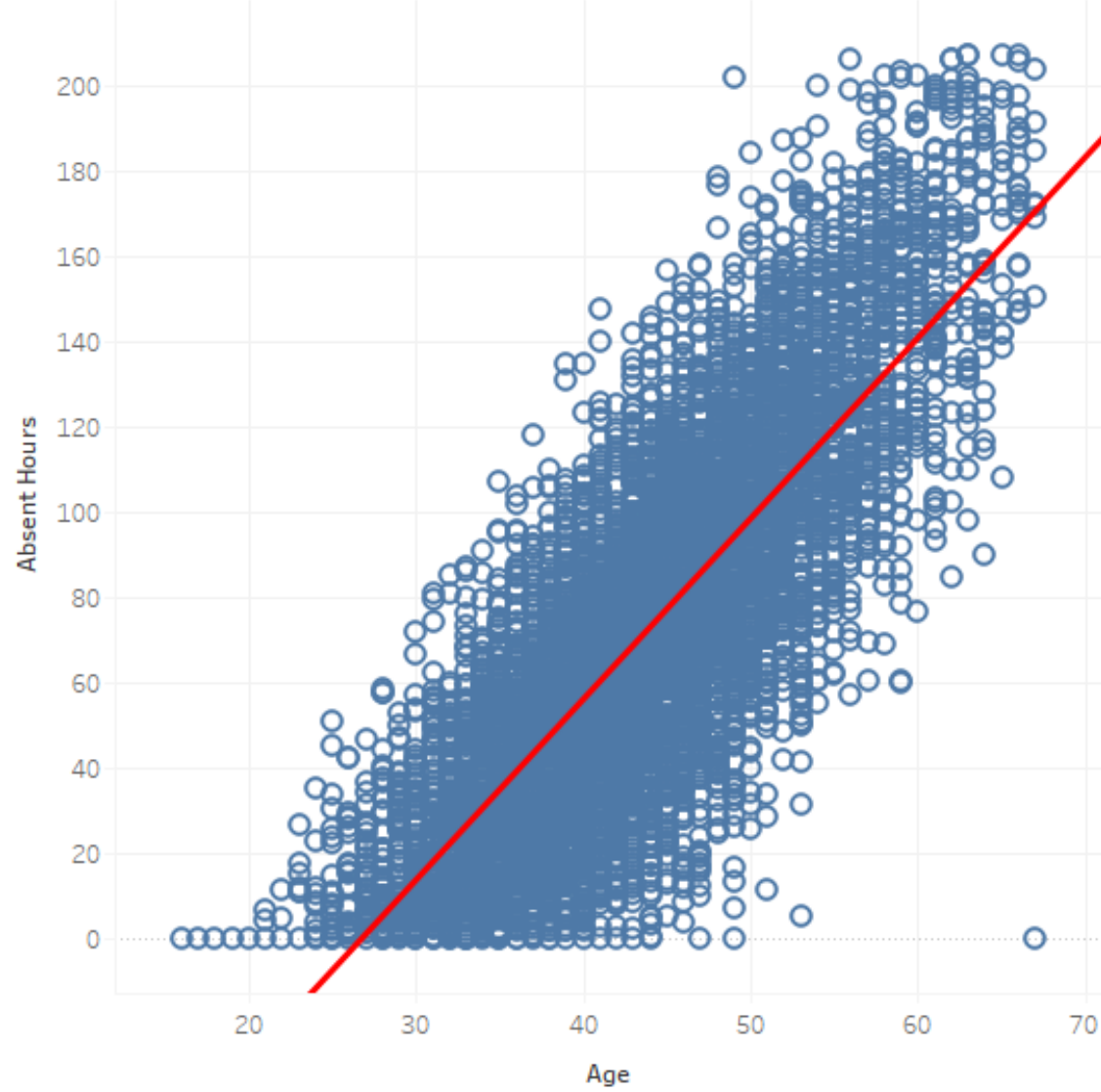


Gender
☒ F
☒ M

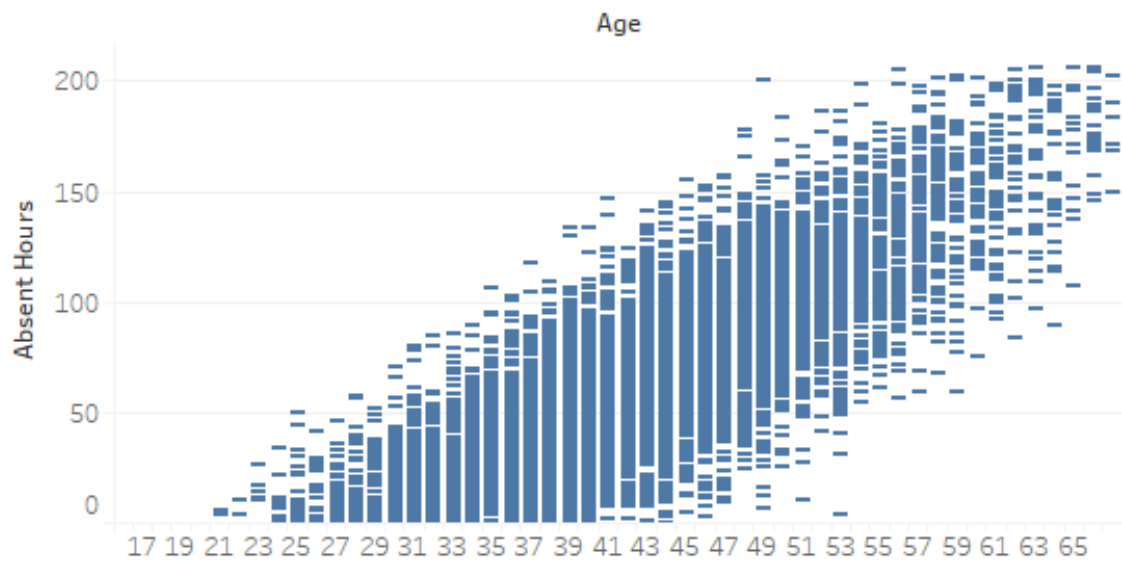
Distribusi Jumlah Karyawan berdasarkan Umur



Trend Lines Simple Linier Regression antara Umur dengan Jam Absen Kerja



Gantt Chart Persebaran Umur Karyawan berdasarkan Jam Absen Kerja





Modelling/Pemodelan

Tujuan Analisis

- Mengetahui korelasi/pengaruh antara umur karyawan dan lama kerja dengan kinerjanya dalam setahun terakhir.
- Memprediksi kinerja karyawan berdasarkan umur dan lama kerja di perusahaan.

*Kinerja karyawan dibuat berdasarkan total jam absen kerja selama setahun terakhir

Metode Analisis

Regression/Prediction

- Simple Linear Regression, $x = \text{age}$, $y = \text{absent_hours}$
- Simple Linear Regression, $x = \text{length_service}$, $y = \text{absent_hours}$
- Multiple Linear Regression

Pemodelan

Independen dan Dependen variabel yang digunakan:

● Independent Variabel

- age
- length_service

● Dependent Variabel

- absent_hours

```
[22] #recall data
      print(data)
```

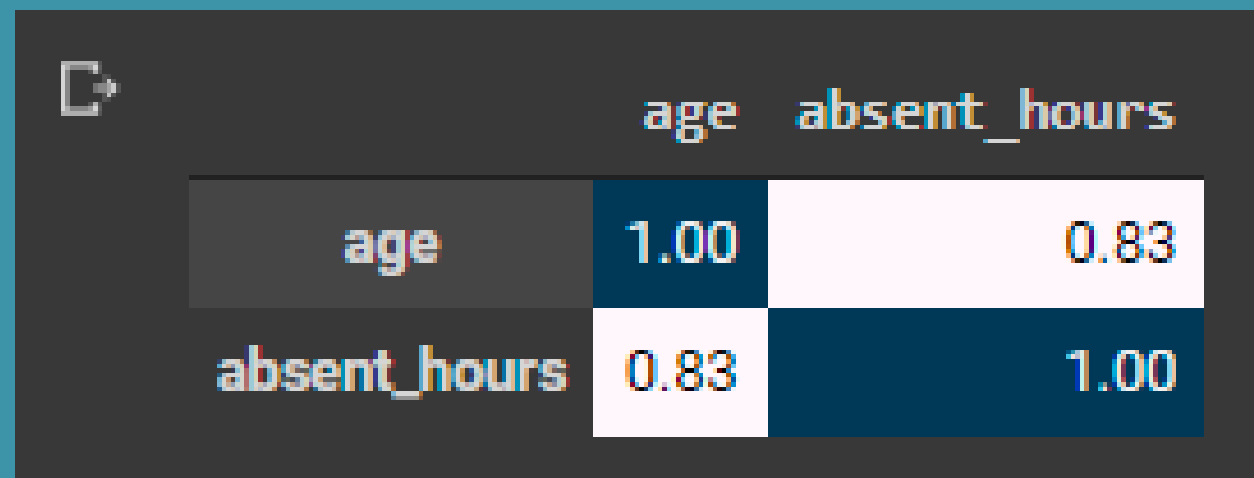
	age	length_service	absent_hours
0	32	6.018478	36.577306
1	40	5.532445	30.165072
2	48	4.389973	83.807798
3	44	3.081736	70.020165
4	35	3.619091	0.000000
...
8331	46	4.838288	93.665111
8332	34	2.427274	0.000000
8333	58	4.009393	176.356940
8334	43	6.154837	60.321917
8335	46	5.174722	112.023389

[8064 rows x 3 columns]

Simple Linier Regression (age, absent_hours)

Modelling dengan metode regresi untuk mengetahui pengaruh umur dengan kinerja karyawan

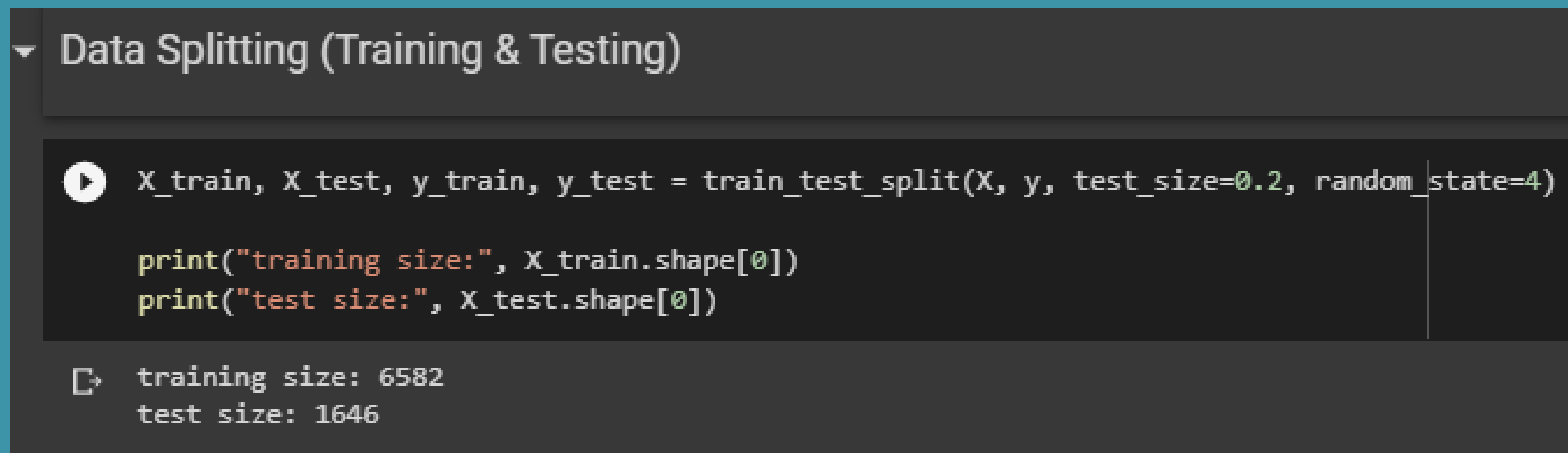
- Correlation variabel independen (age) dengan variabel dependen (absent_hours)



A screenshot of a Jupyter Notebook cell displaying a correlation matrix. The matrix is a 2x2 grid with 'age' and 'absent_hours' as both rows and columns. The diagonal elements are 1.00, and the off-diagonal element is 0.83, indicating a strong positive correlation.

	age	absent_hours
age	1.00	0.83
absent_hours	0.83	1.00

- Data Splitting, memisahkan data training dan testing untuk pemodelan



A screenshot of a Jupyter Notebook cell showing the process of data splitting. The top part contains the code to split the data into training and testing sets using `train_test_split`. The bottom part shows the output of the code, indicating the training size is 6582 and the test size is 1646.

```
▼ Data Splitting (Training & Testing)
```

```
▶ X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=4)

print("training size:", X_train.shape[0])
print("test size:", X_test.shape[0])
```

```
▶ training size: 6582
test size: 1646
```


Simple Linier Regression (age, absent_hours)

- Mencari accuracy score untuk data training

▼ Simple Linear Regression (age)

```
[ ] lin_reg = LinearRegression()  
    #train the model menggunakan training data yang sudah displit.  
    lin_reg.fit(X_train, y_train)
```

```
LinearRegression()
```

```
[ ] print(lin_reg.coef_)  
    print(lin_reg.intercept_)
```

```
[4.11261615]  
-110.14667486484669
```

```
[ ] lin_reg.score(X_test, y_test)
```

```
0.6792051259010135
```

Model mendapatkan accuracy score sekitar 67.9%. Cukup baik untuk iterasi pertama.

Simple Linier Regression (age, absent_hours)

- Menampilkan prediksi jam absen karyawan.

```
[30] y_pred = lin_reg.predict(X_test)
```

```
dataframe = pd.DataFrame({'Data sebenarnya': y_test, 'Data prediksi': y_pred })
```

	Data sebenarnya	Data prediksi
0	12.385091	54.357971
1	105.483054	46.132739
2	0.000000	-23.781736
3	72.204766	62.583203
4	129.633059	79.033668
...
1641	109.181117	91.371516
1642	7.973059	33.794890
1643	5.132366	42.020123
1644	115.275300	120.159829
1645	63.154332	70.808436

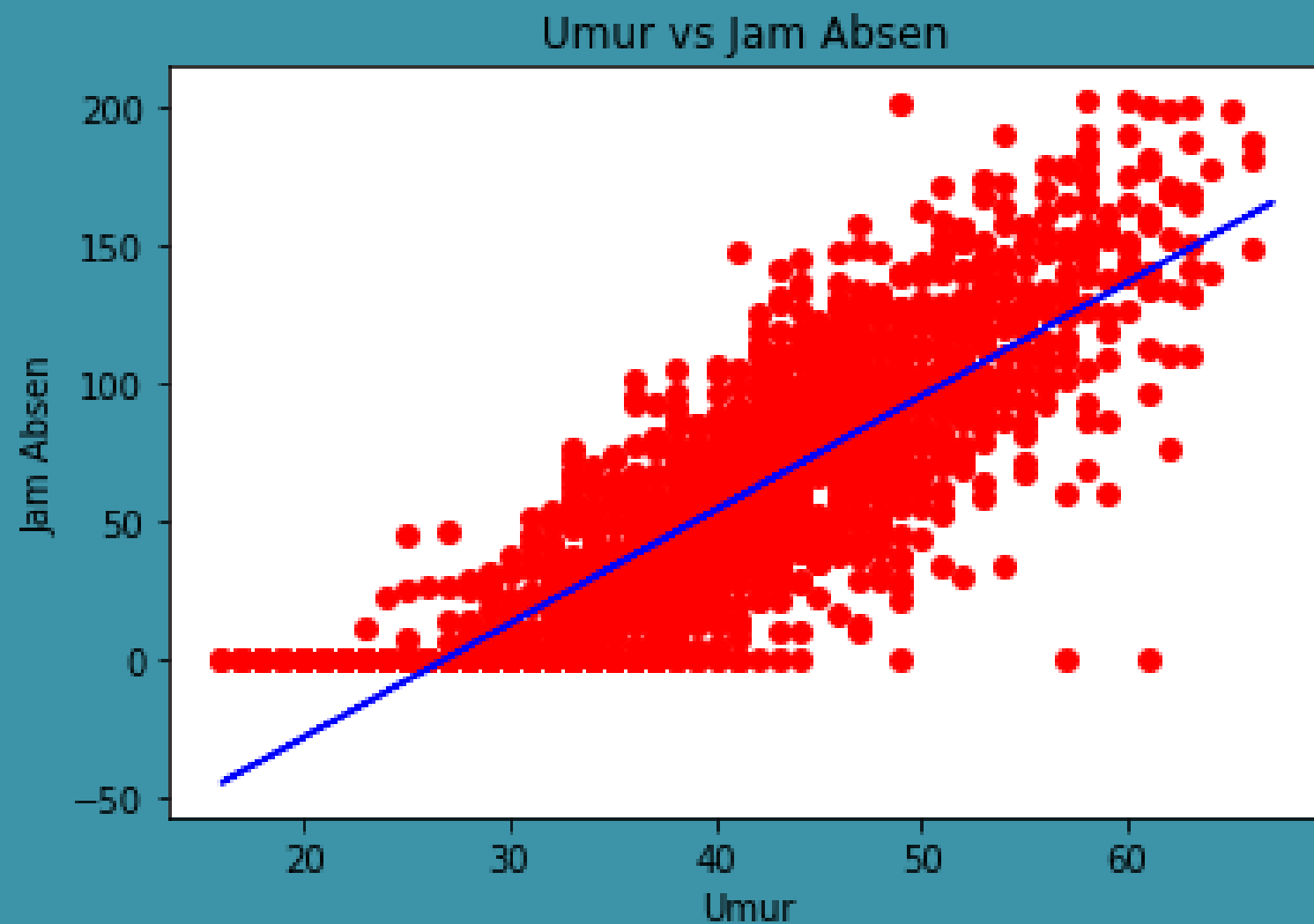
1646 rows x 2 columns

Hasil prediksi jam absen karyawan berdasarkan umur secara otomatis dengan perhitungan linier regresi sederhana

Simple Linier Regression (age, absent_hours)

- Visualisasi scatter plot hubungan antara umur dan jam absen karyawan

```
plt.scatter(X_test, y_test, color = 'red')  
plt.plot(X_train, lin_reg.predict(X_train), color = 'blue')  
plt.title('Umur vs Jam Absen')  
plt.xlabel('Umur')  
plt.ylabel('Jam Absen')  
plt.show()
```



Simple Linier Regression (length_service, absent_hours)

Modelling dengan metode regresi untuk mengetahui pengaruh lama kerja dengan kinerja karyawan

- Data Splitting, memisahkan data training dan testing untuk pemodelan

```
[ ] X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=4)

print("training size:", X_train.shape[0])
print("test size:", X_test.shape[0])

training size: 6528
test size: 1633
```

- Mencari accuracy score untuk data training

```
[ ] lin_reg.score(X_test, y_test)

-4.393830542381849e-05
```

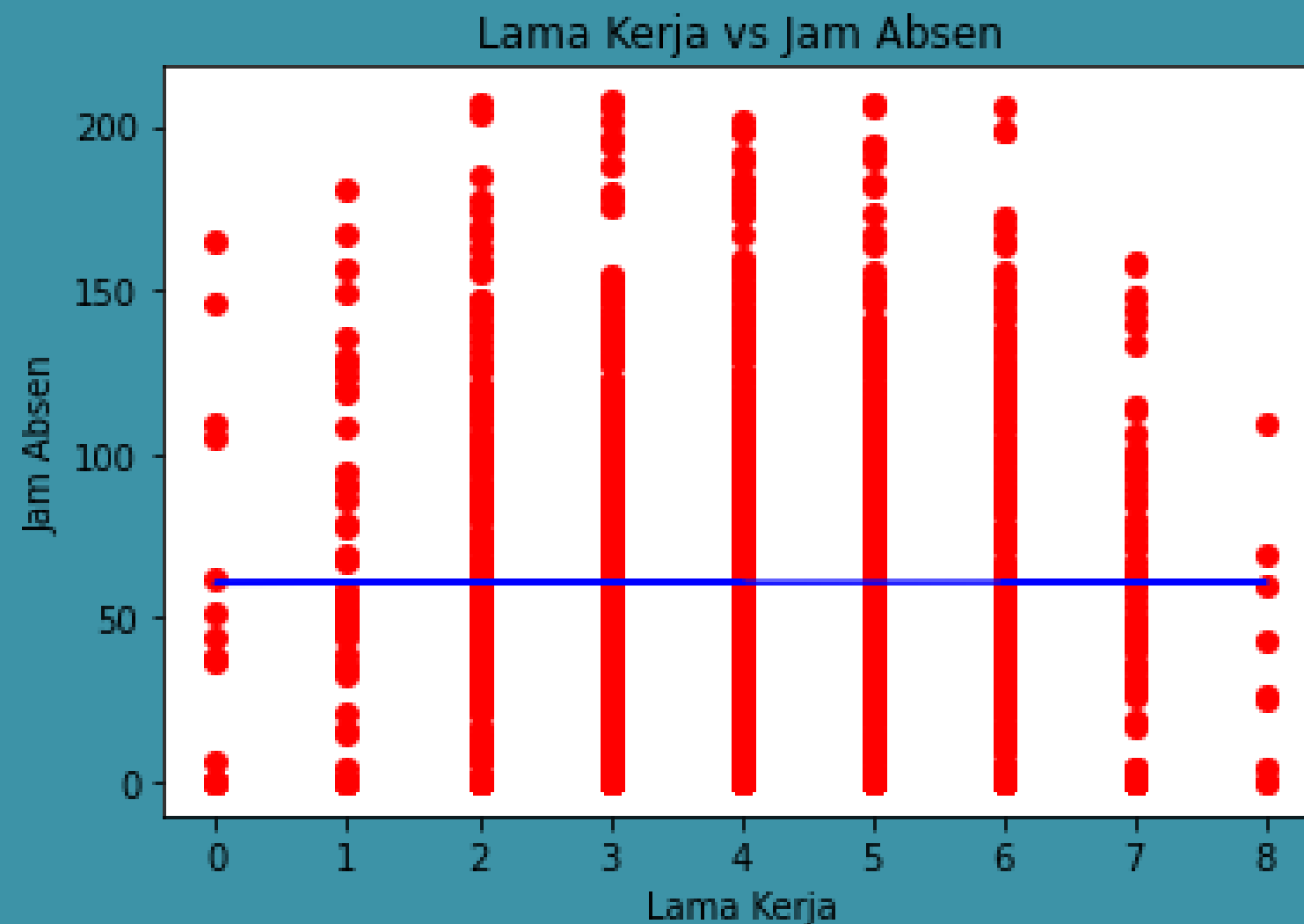
Model mendapatkan accuracy score negatif. Seperti hasil correlationnya, regresi pada kedua variabel ini tidak saling berpengaruh.

Simple Linier Regression (age, absent_hours)

- Visualisasi scatter plot hubungan antara umur dan jam absen karyawan



```
plt.scatter(X_test, y_test, color='red')  
plt.plot(X_train, lin_reg.predict(X_train), color='blue')  
plt.title('Lama Kerja vs Jam Absen')  
plt.xlabel('Lama Kerja')  
plt.ylabel('Jam Absen')  
plt.show()
```



Multiple Linier Regression

Modelling dengan metode multiple regression untuk memprediksi kinerja karyawan berdasarkan umur dan lama kerja di perusahaan.

- Memakai model linear regression menggunakan variabel 'age'

```
✓ [23] X = data[['age']]  
    Y = data['absent_hours']  
    lm = LinearRegression()  
    lm.fit(X,Y)  
    lm.score(X, Y)  
  
0.7047276229110401
```

- Model linear regression untuk memprediksi 'absent_hours' dengan menggunakan variabel 'age' dan 'length_service'

```
✓ [25] features = ["age", "length_service"]  
  
✓ [26] x=data[features]  
    y=data.absent_hours  
    lr.fit(x,y)  
    lr.score(x,y)  
  
0.7047542292874074
```

Multiple Linier Regression

- Memakai list of tuples 'scale', 'polynomial', dan 'model' untuk membuat pipeline yang memprediksi absent_hours

```
[28] Input=[('scale',StandardScaler()),('polynomial', PolynomialFeatures(include_bias=False)),('model',LinearRegression())]

[29] x=data[features]
      y=data.absent_hours
      pipe=Pipeline(Input)
      pipe.fit(x,y)
      pipe.score(x,y)

0.718499628157613
```

- Data splitting untuk multiple regresi

```
[33] features =["age", "length_service"]
      X = data[features]
      Y = data['absent_hours']

      x_train, x_test, y_train, y_test = train_test_split(X, Y, test_size=0.2, random_state=4)

      print("training size:", x_train.shape[0])
      print("test size:", x_test.shape[0])

training size: 6451
test size: 1613
```

Multiple Linier Regression

- Linear least squares dengan l2 regularization/Ridge Regression or Tikhonov regularization
- Transformasi second order polynomial pada training data dan testing data. Membuat Ridge regression object menggunakan training data. Regularisation parameter = 0.1

```
[34] from sklearn.linear_model import Ridge

[35] rm=Ridge(alpha=0.1)
     rm.fit(x_train,y_train)
     rm.score(x_test,y_test)

0.7082041891638853

▶ pr=PolynomialFeatures(degree=2) #second order polynomial
  x_train_pr=pr.fit_transform(x_train) #train data
  x_test_pr=pr.fit_transform(x_test) #test data

  rr=Ridge(alpha=0.1) #Regularization strength; must be a positive float. Regularization in
  rr.fit(x_train_pr,y_train)
  rr.score(x_test_pr,y_test)

0.7206598838099558
```


Multiple Linier Regression

- Membuat model prediksi

```
#Pertama, buat variabel x dan y.  
x = data.drop(columns='absent_hours')  
y = data['absent_hours']  
#Kedua, split data menjadi training and testing dengan porsi 80:20.  
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=4)  
#Ketiga, membuat object linear regresi.  
lin_reg = LinearRegression()  
#Keempat, train the model menggunakan training data yang sudah displit.  
lin_reg.fit(x_train, y_train)
```

- Menghitung accuracy score multiple regression

```
lin_reg.score(x_test, y_test)
```

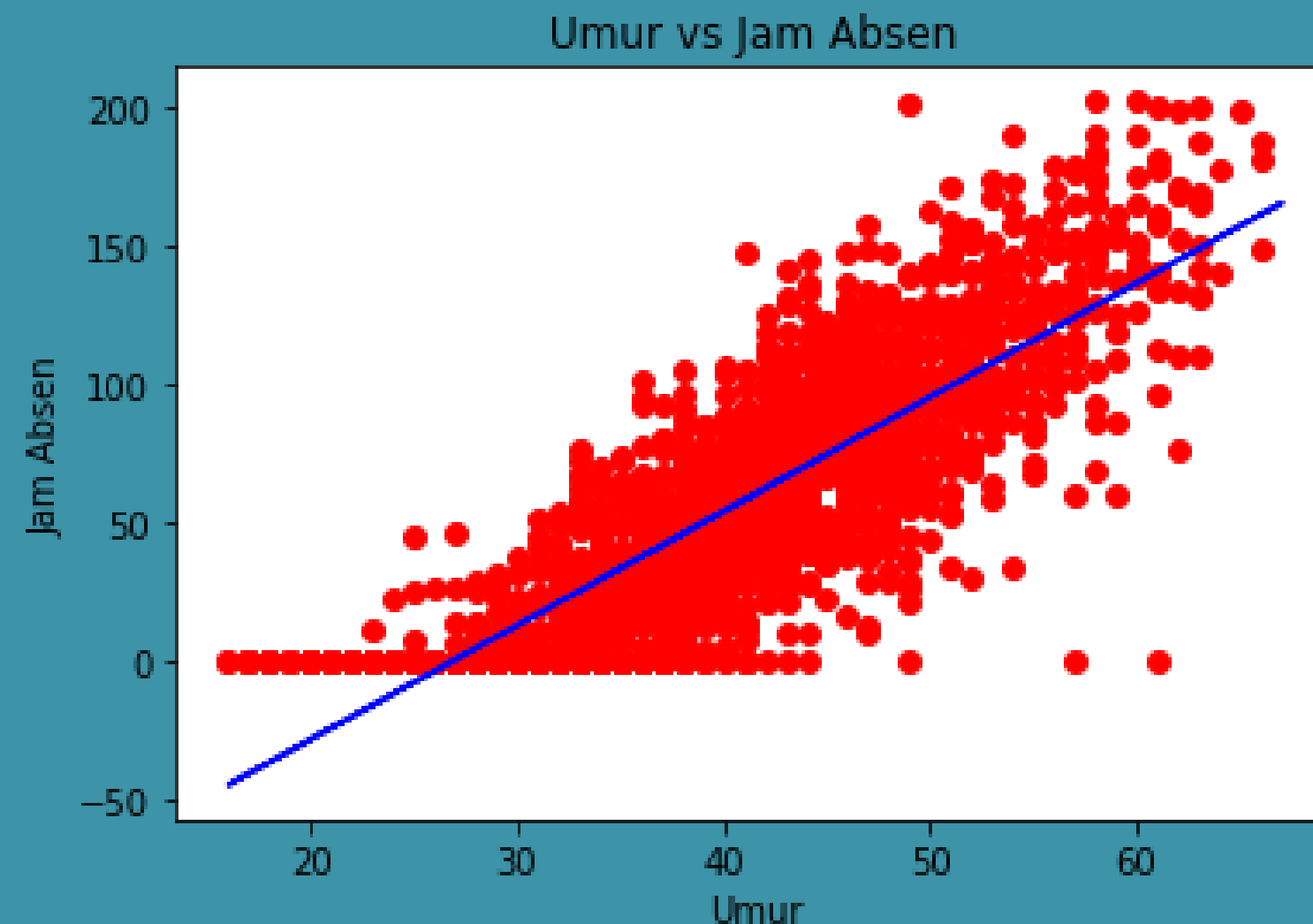
```
0.7082041878375548
```

Multiple Linier Regression

- Mencoba menghitung prediksi dengan menginput nilai age = '60' dan length_service = 10

```
✓ lin_reg.predict([[60, 10.00]])  
array([138.18404438])
```

- Melihat visualisasi scatter plot hubungan antara umur dan jam absen karyawan untuk mencocokkan prediksi.





Hasil Analisis



DEPARTEMEN Human Resource

PEMBAHASAN

1. Umur memiliki pengaruh terhadap kinerja karyawan dimana berdasarkan accuracy score dan visualisasi yang didapat trend lines regression memiliki tren garis naik dan akurasi skor yang cukup akurat di angka sekitar 70%, baik dari hasil pemodelan simple linier, regresi polinomial, maupun regresi ridge.
2. Sedangkan lama kerja berdasarkan accuracy score dan visualisasi yang didapat terlihat distribusi datanya merata dan tidak terlihat kenaikan atau penurunan kinerja karyawan berdasarkan lama kerja.

KESIMPULAN

Variabel umur menjadi fokus dalam pemecahan masalah kinerja karyawan karena insight yang didapat adalah ketika angka umur karyawan semakin tinggi/tua maka jumlah jam absen kerjanya semakin tinggi juga, sehingga disimpulkan umur karyawan memengaruhi kinerjanya. Insight ini bisa menjadi acuan perusahaan dalam mengategorikan karyawan dengan kinerja bagus atau kurang berdasarkan umur dan juga bisa digunakan untuk rekomendasi perusahaan dalam pengelolaan SDM baik untuk rekrutmen, layoff/pengurangan, dan pembaruan requirement dalam merekrut dan menentukan kontrak kerja karyawan di rentang umur produktif.

**Thanks for Looking my
Portfolio :)**