Pacmann

Salary Prediction using Regression Model

Statistics for Business

Diana Darapuspa - ADS Batch 14

Outline

- Introduction
- Dataset
- Statistical Test
- Regression Model
- Conclusion and Recommendation
- References

Introduction

Introduction

Ada satu set data yang memperlihatkan pendapatan seseorang berdasarkan usia, jenis kelamin, tingkat pendidikan, jabatan, dan lama pengalaman kerja. Dalam analisis dataset ini, penulis berkeinginan untuk menilai dampak dari faktor-faktor tersebut terhadap pendapatan dan melakukan prediksi terkait gaji seseorang.

Hal yang akan dilakukan dalam permodelan ini adalah:

1. Menguji pengaruh gender terhadap gaji
2. Memprediksi gaji dari lama pengalaman kerja dengan menggunakan model regresi
3. Memprediksi gaji berdasarkan faktor usia, jenis kelamin, tingkat pendidikan, dan lama pengalaman kerja menggunakan model regresi

Dataset

Dataset

- Datasat yang digunakan bersumber dari Kaggle.com

Dataset ini berisi informasi mengenai gaji karyawan di sebuah perusahaan. Setiap baris mewakili karyawan yang berbeda, dan kolom-kolomnya berisi informasi seperti usia, jenis kelamin, tingkat pendidikan, jabatan, tahun pengalaman, dan gaji.

	Age	Gender	Education_level	Job Title	Years_of_Experience	Salary
0	32.0	Male	Bachelor's	Software Engineer	5.0	90000.0
1	28.0	Female	Master's	Data Analyst	3.0	65000.0
2	45.0	Male	PhD	Senior Manager	15.0	150000.0
3	36.0	Female	Bachelor's	Sales Associate	7.0	60000.0
4	52.0	Male	Master's	Director	20.0	200000.0

Dataset

Usia: Kolom ini mewakili usia setiap karyawan dalam tahun. Nilai dalam kolom ini berupa angka.

Jenis Kelamin: Kolom ini berisi jenis kelamin dari setiap karyawan, yang dapat berupa laki-laki atau perempuan. Nilai dalam kolom ini bersifat kategorikal.

Tingkat Pendidikan: Kolom ini berisi tingkat pendidikan dari setiap karyawan, yang dapat berupa S1, S2, atau S3. Nilai dalam kolom ini bersifat kategorikal.

Tahun Pengalaman: Kolom ini menunjukkan jumlah tahun pengalaman kerja setiap karyawan. Nilai dalam kolom ini adalah angka.

Gaji: Kolom ini mewakili gaji tahunan setiap karyawan dalam dolar AS. Nilai dalam kolom ini adalah angka dan dapat bervariasi tergantung pada faktor-faktor seperti jabatan, tahun pengalaman, dan tingkat pendidikan.

Dataset

- Kolom Jabatan tidak akan digunakan pada permodelan karena terlalu bervariasi
- Data gender merupakan data kategorikal yang akan diubah menjadi data numerik
  Male = 0 dan Female = 1
- Data tingkat pendidikan (Education Level) merupakan data kategorikal yang akan diubah menjadi data numerik
  Bachelor's = 0 ; Master's = 1 ; dan PhD = 2

```
# Use LabelEncoder to convert the smoker variable into numeric
from sklearn.preprocessing import LabelEncoder

# Mapping
gender_mapping = {"Male": 0, "Female": 1}
edu_mapping = {"Bachelor's":0 , "Master's":1 , "PhD":2}

# Create LabelEncoder Object and Transform the Age and education variable
df_salary["Gender"] = LabelEncoder().fit_transform(df_salary["Gender"].map(gender_mapping))
df_salary["Education_level"] = LabelEncoder().fit_transform(df_salary["Education_level"].map(edu_mapping))

```

	Age	Gender	Education_level	Years_of_Experience	Salary
0	32.0	0	0	5.0	90000.0
1	28.0	1	1	3.0	65000.0
2	45.0	0	2	15.0	150000.0
3	36.0	1	0	7.0	60000.0
4	52.0	0	1	20.0	200000.0

Dataset - Data Numerik

- Data numerik terdiri dari kolom Age, Years_of_Experience dan Salary
- Korelasi antara Data numerik

  Korelasi antara usia, lama pengalaman kerja, dan gaji memiliki hasil positif dan berkorelasi kuat

	Age	Years_of_Experience	Salary
Age	1.000000	0.979192	0.916543
Years_of_Experience	0.979192	1.000000	0.924455
Salary	0.916543	0.924455	1.000000

Dataset - Data Kategorik

Data kategorik terdiri dari kolom gender dan education level

```
1 df_salary["Gender"].value_counts()
```

```
Gender
Male      170
Female    154
Name: count, dtype: int64
```

```
1 df_salary["Education_level"].value_counts()
```

```
Education_level
Bachelor's    191
Master's       91
PhD            42
Name: count, dtype: int64
```

Perbandingan gaji berdasarkan data kategorik

Perbandingan Salary based on Education Level

```
1 df_salary.groupby("Education_level")["Salary"].mean()
```

```
Education_level
Bachelor's     73902.356021
Master's      127912.087912
PhD           158095.238095
Name: Salary, dtype: float64
```

Perbandingan Salary based on Gender

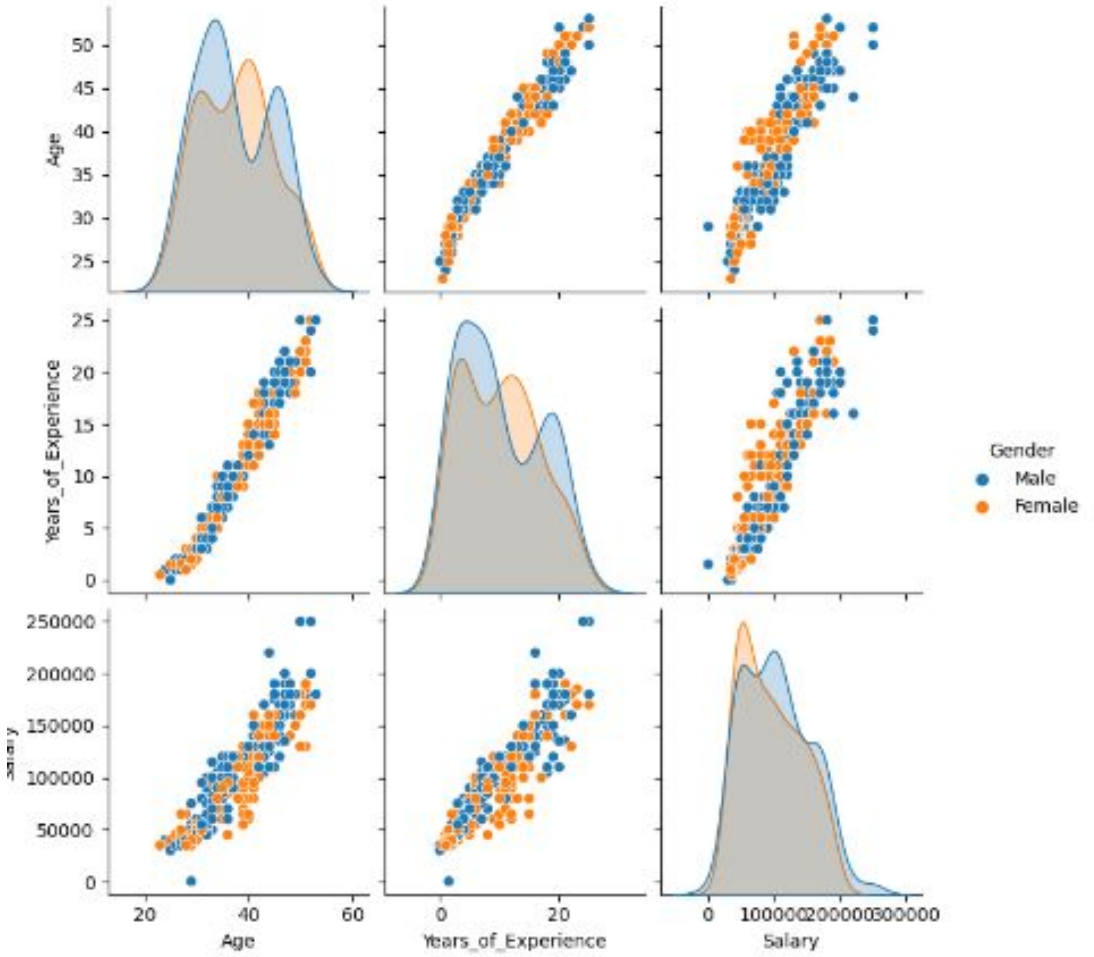
```
1 df_salary.groupby("Gender")["Salary"].mean()
```

```
Gender
Female     96136.363636
Male      103472.647059
Name: Salary, dtype: float64
```

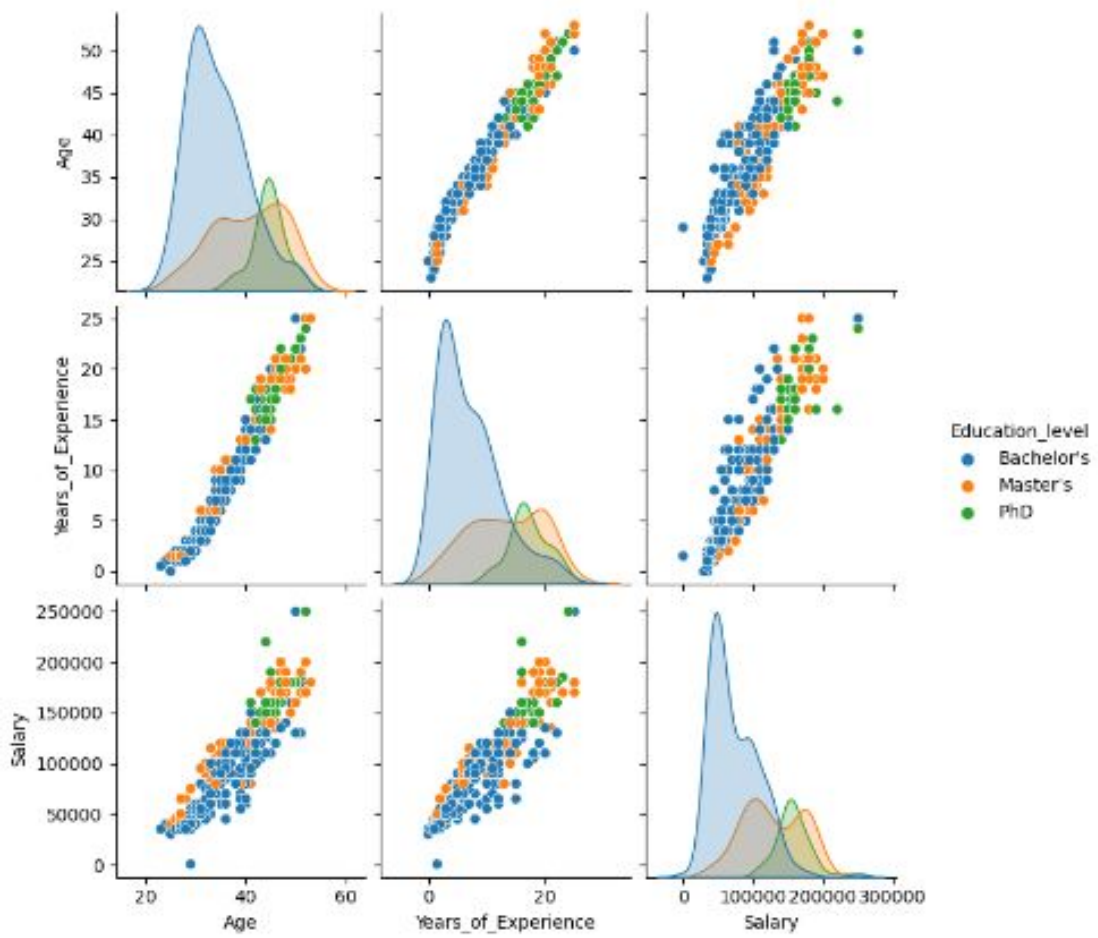
- Rata-rata gaji laki-laki lebih tinggi daripada perempuan
- Rata-rata gaji seseorang yang memiliki tingkat pendidikan lebih tinggi dari Bachelor's cenderung lebih tinggi

Visualisai Data Numerik terhadap Data Kategorik (Gender)

- Semakin bertambah usia seseorang, gaji yang didapat juga semakin tinggi
- Semakin lama pengalaman kerja seseorang, gaji yang didapat juga semakin tinggi
- Variabel jenis kelamin tidak memberikan pola apapun pada setiap variabel numerik terhadap gaji



Visualisasi Data Numerik terhadap Data Kategorik (Education)



- Kenaikan tingkat pendidikan berhubungan positif dengan peningkatan besaran gaji seseorang.
- Seseorang yang lebih tua cenderung memiliki tingkat pendidikan yang lebih tinggi.
- Orang dengan tingkat pendidikan tinggi juga cenderung memiliki pengalaman kerja yang lebih lama.

Uji Statistik

Uji Statistik

Penulis ingin mengetahui apakah gender memengaruhi besarnya gaji yang didapatkan pegawai.
Pada dataset terdapat dua jenis kelamin yaitu Male (a) dan Female (b).

Penulis akan menguji apakah rata-rata gaji yang didapatkan laki-laki lebih tinggi daripada perempuan.
Pengujian dilakukan dengan tingkat signifikansi = 5 %

H_0: Rata-rata gaji laki-laki sama dengan dari rata-rata gaji perempuan.

$$H_0 : \mu_A = \mu_B$$

H_1: Rata-rata gaji laki-laki lebih besar dari rata-rata gaji perempuan.

$$H_1 : \mu_A > \mu_B$$

Uji Statistik

Uji statistik dilakukan dengan menggunakan t-test, karena standard deviasi populasi tidak diketahui.

Uji Variansi

Varians kelompok male dan female tidak sama

```
#Salary Laki-Laki
df_male = df_salary[df_salary["Gender"]=="Male"]["Salary"].values

#Salary Perempuan
df_female = df_salary[df_salary["Gender"]=="Female"]["Salary"].values

#Varianse
np.var(df_male), np.var(df_female)
```

```
(2571353207.6989617, 2097896989.374262)
```

Uji Statistik

Uji statistik dilakukan dengan menggunakan t-test, karena standard deviasi populasi tidak diketahui.

t-test

Dari hasil t-test yang dilakukan didapat bahwa
t-value = 1,346
p-value = 0.08

Karena nilai p-value > tingkat signifikansi
p-value > significance_level
0.08 > 0.05

Maka hasil uji gagal menolak null hypotesis. Dimana dengan tingkat signifikansi sebesar 5% belum ada cukup bukti bahwa rata-rata gaji laki-laki lebih besar dari gaji perempuan

```
result = stats.ttest_ind(a = df_male,
                         b = df_female,
                         equal_var = False,
                         alternative = "greater")
```

```
# Menentukan p-value
result.pvalue
```

```
0.08675461782037655
```

```
result.statistic
```

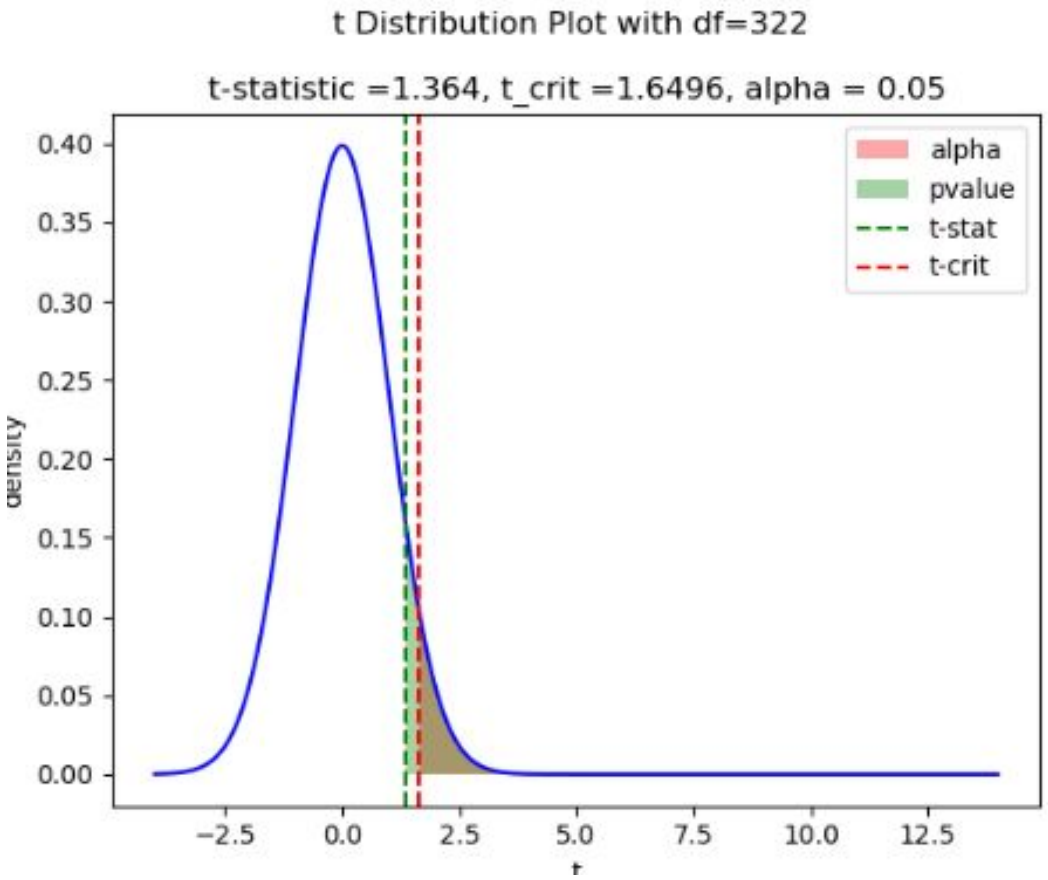
```
1.364034982496829
```

```
# menetukan aturan keputusan
if result.pvalue < significance_level:
    print("Reject the null hypothesis")
else:
    print("Failed to reject the Null hypothesis")
```

```
Failed to reject the Null hypothesis
```

Uji Statistik

Kurva Distribusi t



Confidence Level

```
1  from statsmodels.stats.weightstats import DescrStatsW, CompareMeans
2
3  cm = CompareMeans(d1 = DescrStatsW(data=df_male),
4                    d2 = DescrStatsW(data=df_female))
5
6  lower, upper = cm.tconfint_diff(alpha=significance_level,
7                                  alternative='two-sided',
8                                  usevar='unequal')
9
10 print("Confidence Interval", ":", "[", lower, upper, "]")
```

```
Confidence Interval : [ -3244.897152030464 17917.463996950246 ]
```

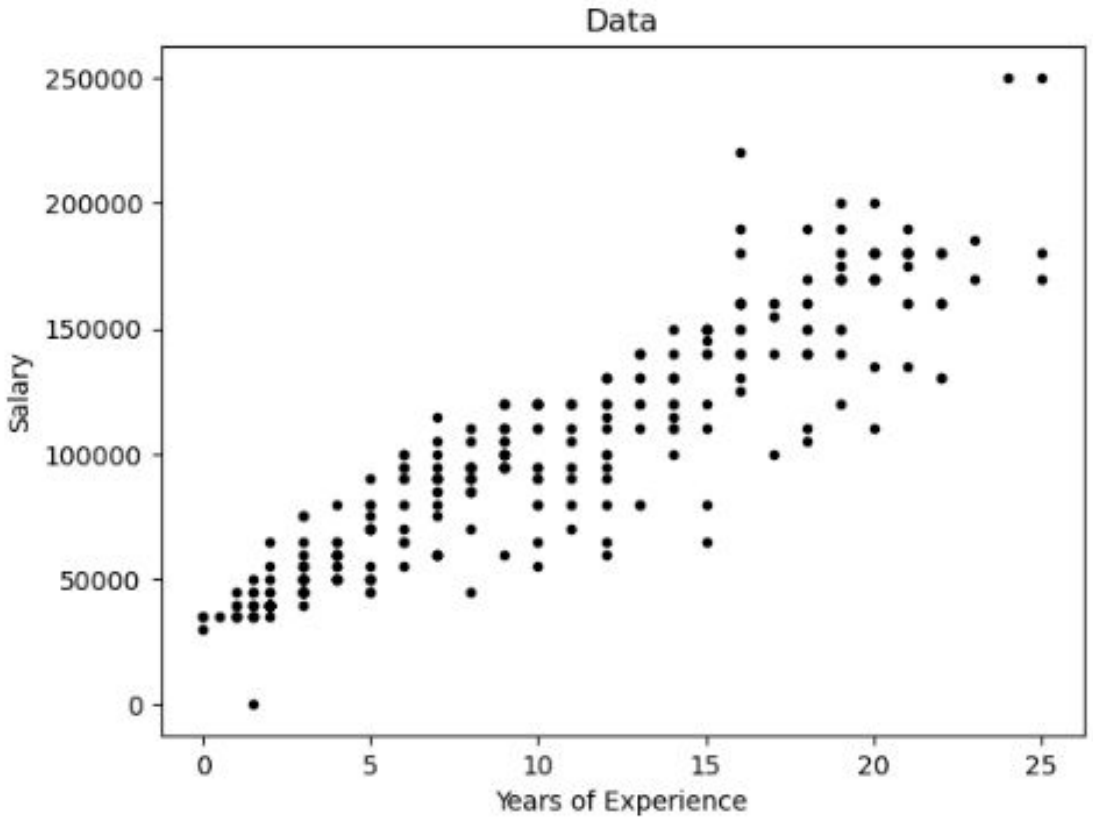
- Dapat disimpulkan bahwa dengan tingkat kepercayaan 95%, penguji yakin bahwa rata-rata gaji laki-laki belum tentu lebih besar dari rata-rata gaji perempuan.
- Dengan tingkat kepercayaan 95%, rata-rata perbedaan gaji memiliki interval di -3245 sampai dengan 17917 dollar

Regression Model

Regression Model : Single Predictor

Grafik berikut menunjukkan hubungan antara pengalaman kerja dan gaji yang didapatkan

Dari grafik berikut, pengalaman kerja dan gaji menunjukkan hubungan positif. Semakin lama pengalaman kerja, variasi gaji yang didapatkan cenderung lebih besar



Regression Model : Single Predictor

```
#Create OLS model object
model = smf.ols("Salary ~ Years_of_Experience", df_salary)

#Fit the model
results_model_salary = model.fit()

#Extract the results (Coefficient and Standard Error) to Dataframe
results_salary = print_coef_std_err(results_model_salary)
results_salary
```

	coef	std err
Intercept	31959.508721	1873.552736
Years_of_Experience	6762.954641	155.446221

```
results_model_salary.rsquared
```

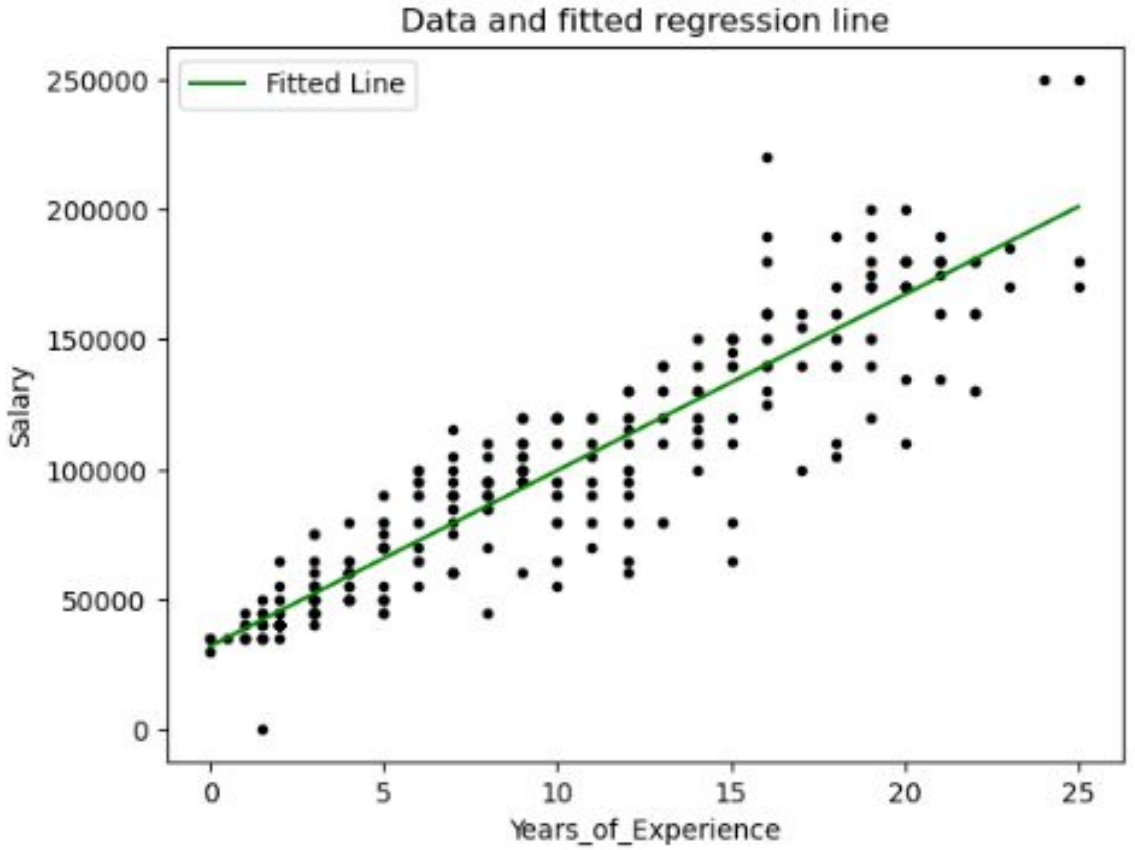
```
0.854616668146078
```

Dari permodelan yang dilakukan didapatkan nilai R-Squared yang cukup baik yaitu bernilai 0.85

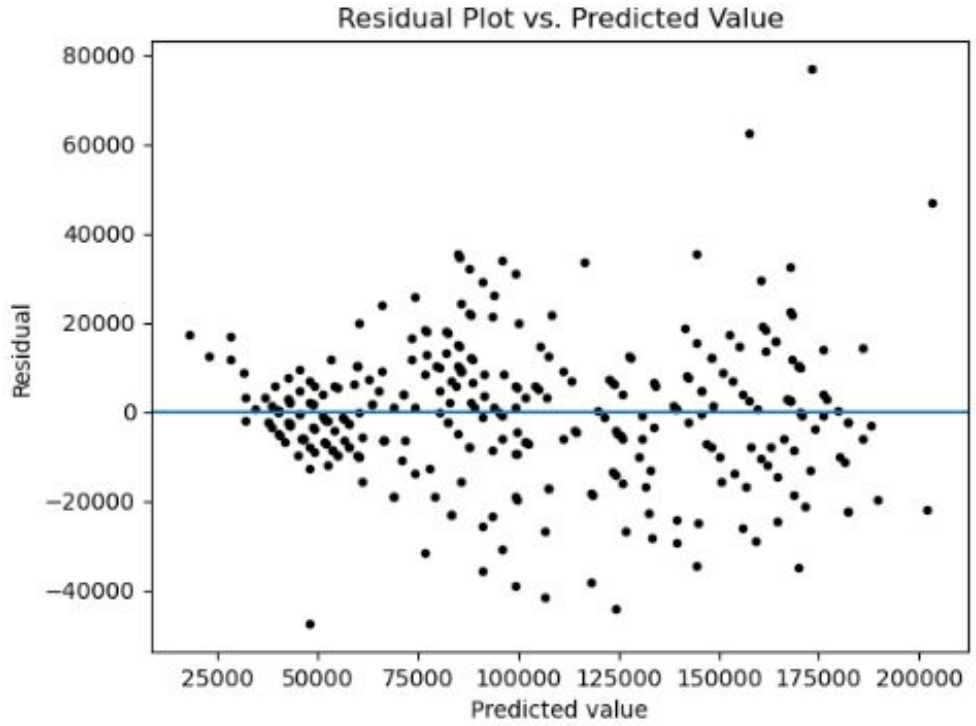
Regression Model : Single Predictor

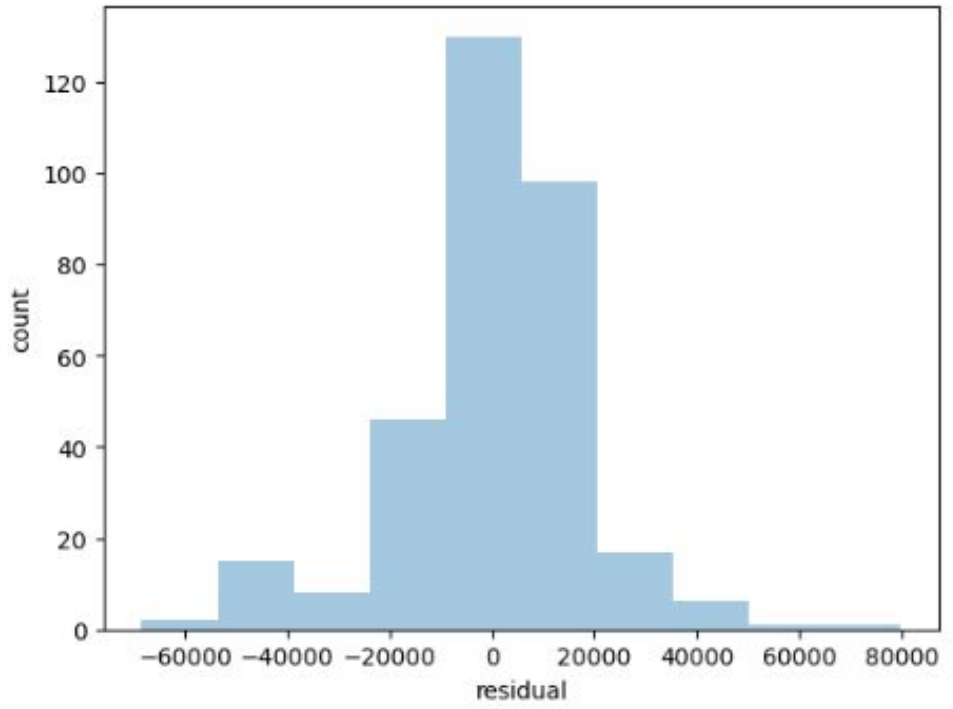
$$\text{Salary} = 31960 + 6763 \times \text{Years of Experience}$$

- Dengan membandingkan dua orang yang memiliki perbedaan 1 tahun di pengalaman kerja, diperkirakan orang yang memiliki pengalaman kerja lebih lama memiliki selisih 6763 lebih besar.
- Untuk orang yang memiliki 0 tahun pengalaman kerja, memiliki perkiraan rata-rata gaji sebesar 31,960



Regression Model : Single Predictor





Residual plot menghasilkan pola yang terlihat jelas, hal ini membuat ketidaksesuaian terlihat meskipun garis regresi menjelaskan lebih dari 85% variansi lama pengalaman kerja

Regression Model : Single Predictor with Log Transformation

```
#Create OLS model object
model = smf.ols("Salary ~ logYoE", df_salary)

#FIt the model
results_logtransform = model.fit()

#Extract the results (Coefficeint and Standard Error) to Dataframe
results_salary_log = print_coef_std_err(results_logtransform)
results_salary_log
```

	coef	std err
Intercept	2321.447994	3284.146736
logYoE	48696.340871	1501.472036

```
results_logtransform.rsquared
```

```
0.7656239539695424
```

Dilakukan permodelan regresi menggunakan transformasi logarithmic pada variabel predictor untuk memprediksi besaran gaji yang diterima pegawai berdasarkan lama pengalaman kerja

Dari hasil permodelan yang dilakukan, nilai **R-Squared = 0.76.**

Hasil R-Squared dari transformasi log lebih rendah dari hasil R-Squared tanpa transformasi log (**R-Squared = 0.85**) . Sehingga untuk permodelan regresi dapat dilakukan tanpa transformasi

Regression Model : Multiple Predictor

Dalam permodelan ini dilakukan regresi menggunakan semua variabel prediktor (usia, jenis kelamin, tingkat pendidikan, dan lama pengalaman kerja). Kemudian ditambah dengan interaksi antara variabel prediktor usia dan pengalaman kerja. Variable kategorikal yang digunakan adalah tingkat pendidikan.

Kemudian dilakukan **Model Evaluation** dengan metode **K-Fold Cross Validation**

```
#Data Splitting Results
fold_train, fold_test = kfold_split(data = df_salary, n_fold=5)
```

```
fold 1, train data rows: 259, test data rows: 65
fold 2, train data rows: 259, test data rows: 65
fold 3, train data rows: 259, test data rows: 65
fold 4, train data rows: 259, test data rows: 65
fold 5, train data rows: 260, test data rows: 64
```

	test_rsquared	folds
0	0.892141	Folds 1
1	0.902729	Folds 2
2	0.912515	Folds 3
3	0.825113	Folds 4
4	0.897267	Folds 5

```
scores_ols_all_pred["test_rsquared"].mean()
```

```
0.8859529642576712
```

Model yang menggunakan semua media memiliki kecocokan yang baik, model ini dapat menjelaskan 88,59% varians gaji.

Regression Model : Multiple Predictor

```
# Create OLS model object
model = smf.ols("Salary ~ Age + Gender + C(Education_level) + Years_of_Expe

# Fit the model
results_model_salary = model.fit()

# Extract the results (Coefficient and Standard Error) to DataFrame
results_salary = print_coef_std_err(results_model_salary)
results_salary
```

	coef	std err
Intercept	-44159.185552	16580.736611
C(Education_level)[T.1]	19574.074815	2257.344892
C(Education_level)[T.2]	26339.473807	3160.610738
Age	3042.039143	611.919060
Gender	-9310.571777	1766.475849
Years_of_Experience	2433.641886	1211.995905
Age:Years_of_Experience	3.452762	21.044653

Setelah dilakukan permodelan dengan multiple predictor didapatkan intercept dan nilai koefision seperti tabel di samping.

Intercept bernilai negatif sehingga kurang bermakna. Karena gaji seseorang dengan pengalaman kerja yang dimulai dari nol tidak mungkin memiliki nilai negatif.

Maka akan dilakukan centering variable usia

Regression Model : Multiple Predictor

```
1 mean_age = df_salary["Age"].mean()
2 mean_age = np.round(mean_age,0)
3 mean_age
```

```
37.0
```

```
1 df_salary["Age"] = df_salary["Age"]-mean_age
2 df_salary.rename(columns = {"Age":"Age_Centered"}, inplace=True)
3 df_salary.head()
```

	Age_Centered	Gender	Education_level	Years_of_Experience	Salary
0	-5.0	0	0	5.0	90000.0
1	-9.0	1	1	3.0	65000.0
2	8.0	0	2	15.0	150000.0
3	-1.0	1	0	7.0	60000.0
4	15.0	0	1	20.0	200000.0

Melakukan Centering Predictor Age

Menggunakan rata-rata usia 37 tahun sebagai acuan. Sehingga data usia akan dihitung jaraknya terhadap rata-rata usia

Regression Model : Multiple Predictor

Setelah melakukan centering predictor, maka kita akan kembali melakukan **Model Evaluation** dengan metode **K-Fold Cross Validation**

	test_rsquared	folds
0	0.849681	Folds 1
1	0.907836	Folds 2
2	0.873470	Folds 3
3	0.938117	Folds 4
4	0.881399	Folds 5

```
1 scores_ols_all_pred["test_rsquared"].mean()
```

```
0.8901007028969223
```

Model yang menggunakan semua media memiliki kecocokan yang baik, model ini dapat menjelaskan 89% varians gaji.

Regression Model : Multiple Predictor

	coef	std err
Intercept	68396.262743	6722.803498
C(Education_level)[T.1]	19574.074815	2257.344892
C(Education_level)[T.2]	26339.473807	3160.610738
Age_Centered	3042.039143	611.919060
Gender	-9310.571777	1766.475849
Years_of_Experience	2561.394070	714.405923
Age_Centered:Years_of_Experience	3.452762	21.044653

Dari hasil permodelan didapatkan nilai intercept dan koefisien variable predictor baru dimana nilai intercept tidak bernilai negatif

Regression Model : Multiple Predictor (Interpretation)

$$\text{Salary for Bachelor's} = 68396 + 3042 \times (\text{Age} - 37) - 9311 \times \text{Gender} + 2561 \times \text{Years\_of\_Experience} + 3 \times (\text{Age} - 37) \times \text{Years\_of\_Experience}$$

$$\text{Salary for Master's} = 68396 + 19574 + 3042 \times (\text{Age} - 37) - 9311 \times \text{Gender} + 2561 \times \text{Years\_of\_Experience} + 3 \times (\text{Age} - 37) \times \text{Years\_of\_Experience}$$

$$\text{Salary for PhD} = 68396 + 26339 + 3042 \times (\text{Age} - 37) - 9311 \times \text{Gender} + 2561 \times \text{Years\_of\_Experience} + 3 \times (\text{Age} - 37) \times \text{Years\_of\_Experience}$$

Interpretasi Intercept

Seorang laki-laki yang berusia 37 tahun dengan tingkat pendiidikan Bachelor’s dan tidak memiliki pengalaman kerja, diperkirakan akan memiliki gaji sebesar 68,396 dollar

Interpretasi Usia

Jika mengamati dua individu dengan jenis kelamin dan tingkat pendidikan yang sama, dengan lama pengalaman kerja = 0, perkiraan gaji seseorang yang memiliki usia lebih 1 tahun dari 37 tahun, diperkirakan lebih tinggi sebesar 3042 dolar dibandingkan dengan individu yang memiliki usia 37 tahun

Regression Model : Multiple Predictor (Interpretation)

$$\text{Salary for Bachelor's} = 68396 + 3042 \times (\text{Age} - 37) - 9311 \times \text{Gender} + 2561 \times \text{Years\_of\_Experience} + 3 \times (\text{Age} - 37) \times \text{Years\_of\_Experience}$$

$$\text{Salary for Master's} = 68396 + 19574 + 3042 \times (\text{Age} - 37) - 9311 \times \text{Gender} + 2561 \times \text{Years\_of\_Experience} + 3 \times (\text{Age} - 37) \times \text{Years\_of\_Experience}$$

$$\text{Salary for PhD} = 68396 + 26339 + 3042 \times (\text{Age} - 37) - 9311 \times \text{Gender} + 2561 \times \text{Years\_of\_Experience} + 3 \times (\text{Age} - 37) \times \text{Years\_of\_Experience}$$

Interpretasi Tingkat Pendidikan

Jika mengamati dua individu dengan usia, jenis kelamin, dan lama pengalaman kerja yang identik, perkiraan gaji seseorang yang memiliki gelar Master's diperkirakan lebih tinggi sebesar 19574 dolar dibandingkan dengan individu yang memiliki gelar Bachelor's.

Kemudian jika mengamati dua individu dengan usia, jenis kelamin, dan lama pengalaman kerja yang identik, perkiraan gaji seseorang yang memiliki gelar PhD diperkirakan lebih tinggi sebesar 26339 dolar dibandingkan dengan individu yang memiliki gelar Bachelor's.

Regression Model : Multiple Predictor (Interpretation)

$$\text{Salary for Bachelor's} = 68396 + 3042 \times (\text{Age} - 37) - 9311 \times \text{Gender} + 2561 \times \text{Years\_of\_Experience} + 3 \times (\text{Age} - 37) \times \text{Years\_of\_Experience}$$

$$\text{Salary for Master's} = 68396 + 19574 + 3042 \times (\text{Age} - 37) - 9311 \times \text{Gender} + 2561 \times \text{Years\_of\_Experience} + 3 \times (\text{Age} - 37) \times \text{Years\_of\_Experience}$$

$$\text{Salary for PhD} = 68396 + 26339 + 3042 \times (\text{Age} - 37) - 9311 \times \text{Gender} + 2561 \times \text{Years\_of\_Experience} + 3 \times (\text{Age} - 37) \times \text{Years\_of\_Experience}$$

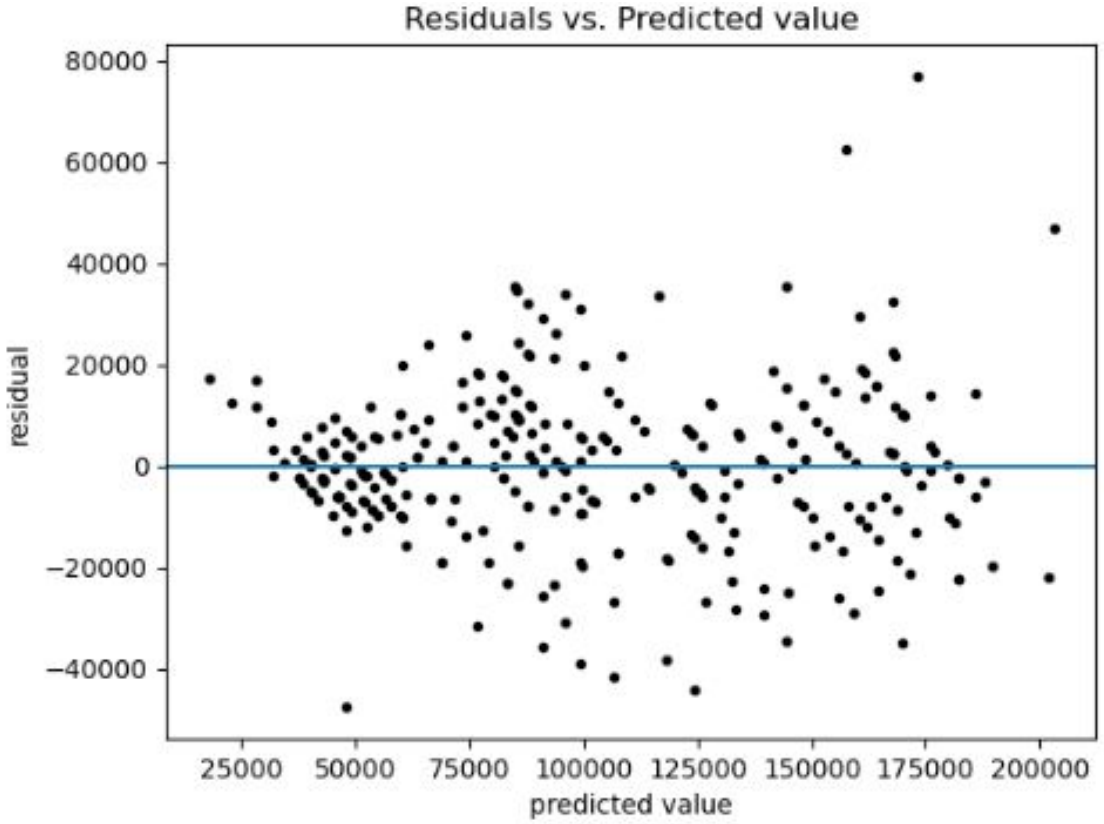
Interpretasi Jenis Kelamin

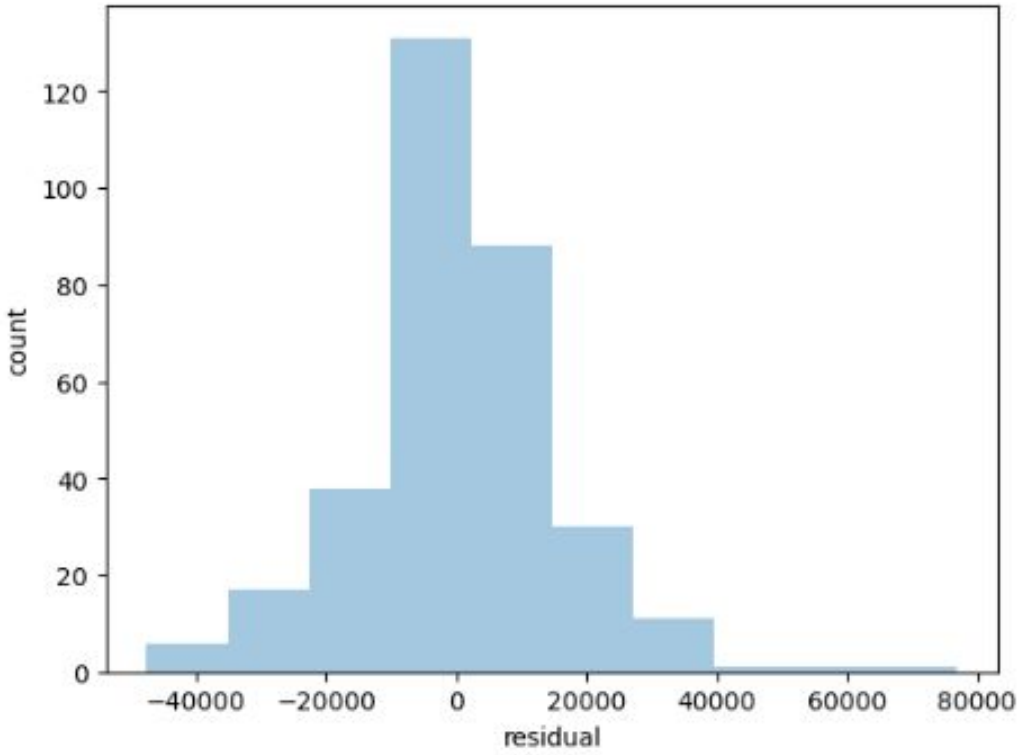
Jika mengamati dua individu yang memiliki usia, lama pengalaman kerja, dan tingkat pendidikan yang sama, perkiraan gaji seorang perempuan lebih sedikit 9311 dollar dibandingkan dengan seorang laki-laki.

Interpretasi Pengalaman Kerja

Jika mengamati dua individu dengan usia = 37 tahun, jenis kelamin, dan tingkat pendidikan yang identik, perkiraan gaji seseorang yang memiliki pengalaman kerja lebih lama 1 tahun diperkirakan lebih tinggi sebesar 2561 dolar.

Regression Model : Multiple Predictor





Conclusion and Recommendations

Conclusion

- Dapat disimpulkan bahwa usia, jenis kelamin, tingkat pendidikan, pengalaman kerja berpengaruh terhadap besaran gaji pegawai.
- Model regresi yang dibangun dengan single predictor (Years_of_Experience) memiliki performance yang cukup bagus yaitu R-Squared = 0.85. Sedangkan jika menggunakan transformasi logaritmic menghasilkan R-Squared = 0.76, dimana hasilnya lebih rendah.
- Model regresi dengan multiple predictor memiliki performa yang lebih baik, nilai R-Squared = 0,89. Dimana telah dilakukan centering pada variabel usia agar memilki interpretasi yang lebih bermakna

Recommendation

- Untuk pengujian lebih lanjut dapat dilakukan percobaan dengan berbagai variasi predictor. Dapat dikelompokkan berdasarkan jabatan / jenis pekerjaan, level-level tertentu dari pengelompokan tingkat pendidikan dan lama pekerjaan sehingga dapat menghasilkan prediksi range gaji berdasarkan kombinasi predictor

Reference

- An introduction to statistical learning by James, G., Witten, D., Hastie, T., & Tibshirani, R
- Statistics for Business : Decision Making and Analysis — Robert Stine and Dean Foster
- Regression and Other Stories by Andrew Gelman, Jennifer Hill, Aki Vehtari
- The Effect: An Introduction to Research Design and Causality. Chapter 13 Huntington-Klein, N. 2021

Thank You