Laporan Tugas Eksplorasi Dataset

"Pendapatan Penduduk Kota San Francisco"

Penjelasan Singkat *Dataset*

Dataset ini menyediakan data gaji penduduk Kota San Francisco dari tahun 2011 sampai 2014, untuk berbagai macam *job title*. Setiap *record* pada *dataset* ini terdiri dari data ID, EmployeeName, JobTitle, BasePay, OvertimePay, OtherPay, Benefits, TotalPay, TotalPayBenefits, Year, Notes, Agency, Status (Fulltime/ Parttime). Pada *dataset* ini terdapat 148.654 *records.*

Dari *dataset* ini, dapat dilakukan analisis mengenai beberapa hal, antara lain:

a. *Trend* rata-rata pendapatan penduduk San Francisco dari tahun ke tahun beserta prediksi untuk tahun-tahun berikutnya
b. Pengelompokan penduduk San Francisco berdasarkan besar pendapatannya (rendah, sedang, dan tinggi)
c. Pengelompokan penduduk San Francisco berdasarkan bidang pekerjaannya (misal: Engineering, Politics, Public Services, Health Care, dll.)

Langkah-Langkah Analisis

Langkah pertama yang dilakukan yaitu melihat *dataset* secara sekilas. Pada *dataset* terdapat beberapa kolom yang menyediakan informasi penting, yaitu *job title, total pay-benefits,* dan *year.* Dari ketiga kolom tersebut dapat dilakukan analisis mengenai rata-rata pendapatan penduduk berdasarkan tahun serta pengelompokan penduduk berdasarkan pendapatan dan bidang pekerjaannya.

Setelah melihat *dataset,* kami menentukan *tools* yang akan digunakan. Kami menggunakan Bahasa R untuk analisis data dan tambahan *library* RTextTools untuk klasifikasi serta *library* Ggplot2 untuk visualisasi data. Kemudian kami melakukan eksplorasi terhadap *tools* tersebut untuk menganalisis data.

Perhitungan Rata-Rata Gaji dan Prediksi

Untuk menghitung rata-rata, kami menggunakan fungsi aggregate() yang mengelompokkan data berdasarkan kolom tertentu dan langsung menghitung rata-rata dari data yang telah dikelompokkan tersebut.

Untuk memprediksi rata-rata pendapatan dua tahun setelahnya (2015 dan 2016), pertama kami memodelkan dengan regresi linear data rata-rata yang telah didapatkan dengan menggunakan fungsi lm(). Setelah itu kami mengecek hasilnya untuk memastikan apakah model linear dapat digunakan untuk melakukan prediksi. Namun karena data yang ada hanya sedikit, hasil regresi secara linear ternyata memiliki *p-value* yang tinggi. Setelah mencoba model lain dan tidak ada yang bisa dengan tepat memodelkan data (karena jumlah data yang sangat sedikit), kami memutuskan tetap memakai model

linear tersebut untuk memprediksi menggunakan fungsi predict.lm(), dengan asumsi bahwa akan terjadi peningkatan gaji untuk tahun-tahun berikutnya.

Pengelompokan Penduduk berdasarkan Gaji

Langkah yang dilakukan adalah sebagai berikut.

1. Ambil kolom gaji total (TotalPayBenefits) dari dataset
2. Pisahkan gaji total tersebut per tahun
3. Buat batasan gaji yang tergolong rendah, sedang, dan tinggi dengan menggunakan fungsi cut(). Batasan ditentukan dengan rata-rata keseluruhan dan standar deviasi setiap tahunnya
4. Akan terbentuk data baru yang berisi tulisan Rendah, Sedang, atau Tinggi. Data tersebut tentunya masih terpisahkan per tahun. Sebelum divisualisasikan, data harus disatukan terlebih dahulu.

Pengelompokan Penduduk berdasarkan Jenis Pekerjaan

Semua langkah – langkah analisis ini termuat dalam *file* "Classification.R". Permasalahan yang diharapkan dapat diselesaikan dalam hal ini adalah pengelompokan atau pengklasifikasian penduduk San Fransisco berdasarkan bidang pekerjaannya.

Ide utamanya adalah membuat *data frame* yang mendaftarkan pasangan pekerjaan dan kelompok pekerjaannya. Kemudian daftar tersebut akan digunakan untuk mengisi kolom bidang yang ditambahkan ke *data frame* utama. Analisis dimulai dengan eksplorasi terhadap teknik pengelompokan yang dibagi menjadi dua, yaitu *supervised* dan *unsupervised*. Sederhananya *supervised* memerlukan sampel dari user untuk digunakan oleh mesin dalam pengelompokan sementara *unsupervised* tidak memerlukan. Tentunya untuk menggunakan teknik tersebut pertama kali akan dilakukan manipulasi data untuk membuang informasi yang tidak diperlukan.

Dikarenakan *unsupervised* tampak lebih mudah pertama dilakukan pendekatan *unsupervised* untuk *dataset* ini. Beberapa teknik yang ditemukan dalam eksplorasi adalah *k-means clustering* dan juga *hierarchical clustering.* Namun *k-means* membutuhkan penggunan untuk secara langsung menentukan jumlah *cluster* yang dalam hal ini menentukan jumlah bidang pekerjaan. Tentunya hal ini sulit dilakukan jika tidak diperiksa *manual* untuk dataset seperti ini. Karena itu dicoba pendekatan *hierarchical clustering*. *Dissimilarity matrix* yang dibutuhkan didapatkan dengan menggunakan *package* stringdist yang menggunakan OSA (Optimal String Alignment) terhadap pekerjaan. Setelah dendogram terbentuk, dari hasil yang ada cukup sulit untuk menentukan di mana sebaiknya dilakukan pemotongan. Selain itu disadari juga kemiripan atau jarak antar *string* yang jauh tidak berarti pekerjaan tersebut memiliki kategori yang berbeda. Menyadari kelemahan ini, penulis kemudian memutuskan untuk mengganti pendekatan lagi, kali ini menjadi *supervised classification*.

Salah satu *package* yang ada dan dapat dimanfaatkan dalam *supervised classification* untuk *text* adalah RTextTools. Penulis kemudian memanfaatkan *package* ini untuk mencoba menyelesaikan persoalan yang diberikan.

Pertama, perlu disiapkan *training data* yang akan digunakan mesin untuk mempelejari pengelompokan. Penulis mempertimbangkan dua alternative yaitu melakukan pemilihan secara acak dan kemudian mengelompokannya atau melakukan pemilihan secara *manual*. Karena dikhawatirkan hasil dari

pemilihan secara acak akan kurang mewakilkan beberapa bidang, maka dilakukan pemilihan secara *manual*. Hasilnya adalah 526 dari 1234 data terpilih menjadi *training data*. *Training data* ini berisi *job* dan juga *category* yang kemudian menjadi sampel untuk mesin. Kemudian dari *list* pekerjaan yang sudah dibersihkan dibuat list baru dengan membuang *training data* sehingga sisanya menjadi *test data*. Tentunya ditambahkan *dummy* dalam kolom *category* pada *test data*.

Berdasarkan daftar pekerjaan yang sudah dibersihkan lagi dibuat juga *document-term matrix*, sebuah *matrix* yang menghitung kemunculan *term* dalam *document* (dalam hal ini *document* adalah *Job*). Dari *matrix* ini kemudian dibuat *container* yang menyimpan *train* dan *test* data yang ada beserta *document-temr matrix*.

Container kemudian dimanfaatkan untuk pembuatan *model* dan *model* dimanfaatkan untuk melakukan prediksi. Terdapat berbagai teknik pemodelan yang digunakan dan dari membandingkan jumlah kategori *others* pada hasil prediksi serta pengecekan *manual* secara singkat, dipilihlah model berdasarkan *Maximum Entropy*.

Hasil prediksi dengan menggunakan model tersebut kemudian diperiksa dan dikoreksi secara *manual*. Setelah selesai hasil prediksi tersebut digabungkan dengan *train data* menjadi sebuah *data frame* yang mendaftarkan dengan lengkap jenis pekerjaan. Sesuai dengan penjelasan pada awal bagian ini, *data frame* ini kemudian digunakan untuk mengisi kolom *category* pada data frame utama. Setelah pengisian tersbut kemudian dilakukan *plotting* frekuensi kemunculan tiap kategori pada *data frame* serta penyimpanan hasil olahan *data frame* utama ke dalam *format* csv.

Hasil Analisis

Script dan kode serta *file* .csv terlampir pada *deliverables.*.

Perhitungan Rata-Rata Gaji dan Prediksi

Dari hasil agregasi didapatkan rata-rata gaji penduduk San Francisco sebagai berikut.

	tahun	gaji
1	2011	71744.1
2	2012	100553.2
3	2013	101440.5
4	2014	100250.9

Hasil regresi linear data diatas adalah sebagai berikut.

```
Call:
lm(formula = gaji ~ tahun, data = ratarata)

Residuals:
    1     2     3     4
-8792 11376  3623 -6207

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -17296060   10229738  -1.691    0.233
tahun            8641       5083   1.700    0.231

Residual standard error: 11370 on 2 degrees of freedom
Multiple R-squared:  0.591,     Adjusted R-squared:  0.3865
F-statistic:  2.89 on 1 and 2 DF,  p-value: 0.2313
```

Hasil prediksi dengan menggunakan regresi linear untuk rata-rata gaji tahun 2015 dan 2016 adalah sebagai berikut.

	tahun	gaji
1	2015	115099.1
2	2016	123739.9

Pengelompokan Penduduk berdasarkan Gaji

Ada 3 golongan gaji di kota San Fransisco, yaitu Rendah, Sedang, dan Tinggi. Dengan perhitungan menggunakan rata-rata dan standar deviasi (Rendah = (0,mean-standar deviasi), Sedang = (mean-standar deviasi, mean+standar deviasi), dan Tinggi = (mean+standar deviasi, nilai maksimal)) didapatkan 32104 data tergolong Rendah, 92393 tergolong Sedang, dan 24157 tergolong Tinggi.

Pengelompokan Penduduk berdasarkan Jenis Pekerjaan

- *Factor* setelah manipulasi: 1234 *Levels*
- *Train* data: 526
- *Test* data: 708
- *Document-Term Matrix*

<<DocumentTermMatrix (documents: 1234, terms: 775)>>
Non-/sparse entries: 3779/952571
Sparsity : 100%
Maximal term length: 25
Weighting : term frequency - inverse document frequency (normalized) (tf-idf)

- Jumlah *Others* pada hasil prediksi

Model	Jumlah Others
TREE	480
MAXENT	180

SVM	383
SLDA	334
BAGGUBG	539
BOOST	362

- Hasil Pengecekan Manual
  Sekitar 183 klasifikasi yang salah dari hasil prediksi menggunakan MAXENT.

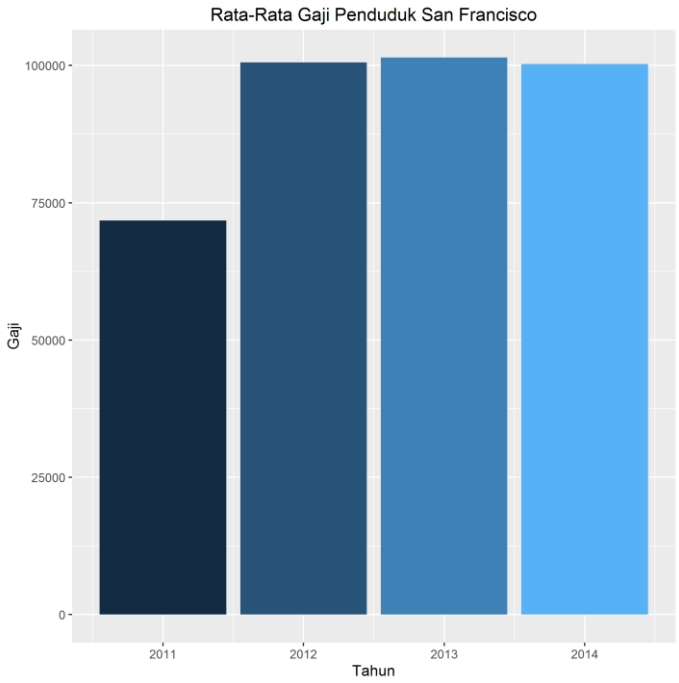
- Daftar lengkap *Job* dan *Category*
  Karena hasil yang sangat panjang, maka hasil tidak ditampilkan dalam dokumen ini namun dapat dilihat dengan menjalankan semua perintah (*Local Setup* disesuaikan) pada Classification.R yang terdapat pada *deliverables*. Hasil tersimpan dalam *variable* bernama complete.

- Frekuensi setiap *Category* dalam *dataset*

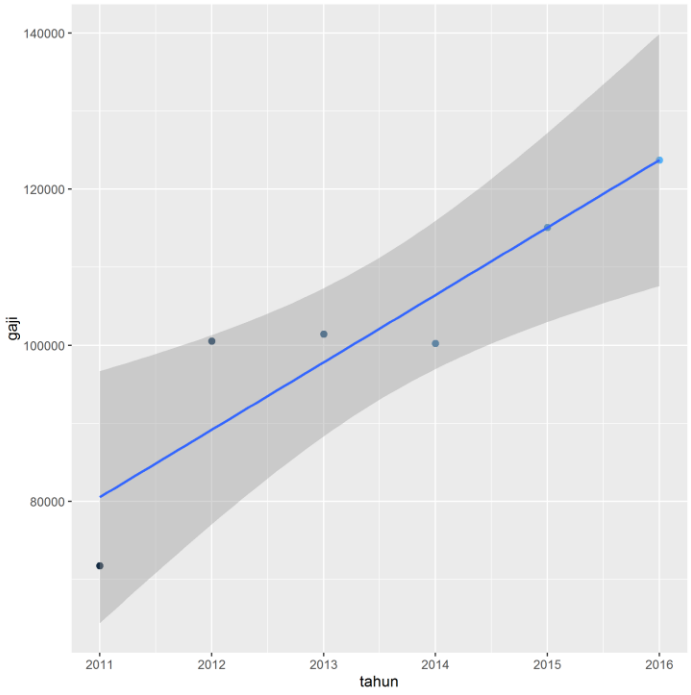
No	Category	Frekuensi Kemunculan
1	Administration	2212
2		2
3	Business	938
4	Communication	553
5	Construction	2143
6	Energy	2712
7	Engineering	5034
8	Environment	2157
9	Finance	2138
10	Healthcare	16329
11	Human Relations	129
12	Human Resources	3903
13	IT	626
14	Law	6092
15	Manufacture	148
16	Marketing	4
17	Media	286
18	Others	61916
19	Properties	286
20	Public Relations	712
21	Public Services	8413
22	Science	2825
23	Security	15585
24	Services	10208
25	Transportation	3303

- Hasil Pengelompokan (Hasil Akhir)
  Karena hasil yang terlalu panjang, maka hasil tidak dimuat dalam dokumen ini namun dapat dilihat pada *file* Classification.csv yang dilampirkan pada *deliverables*.

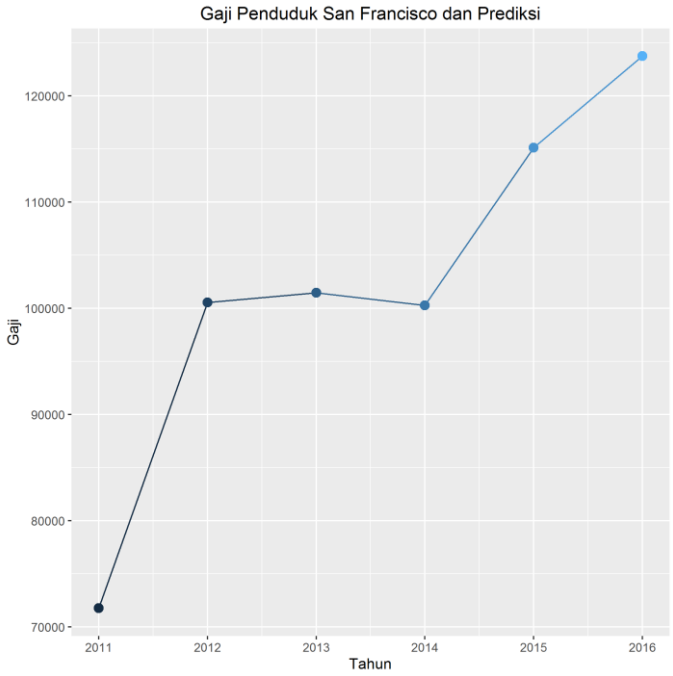
Visualisasi Hasil Analisis



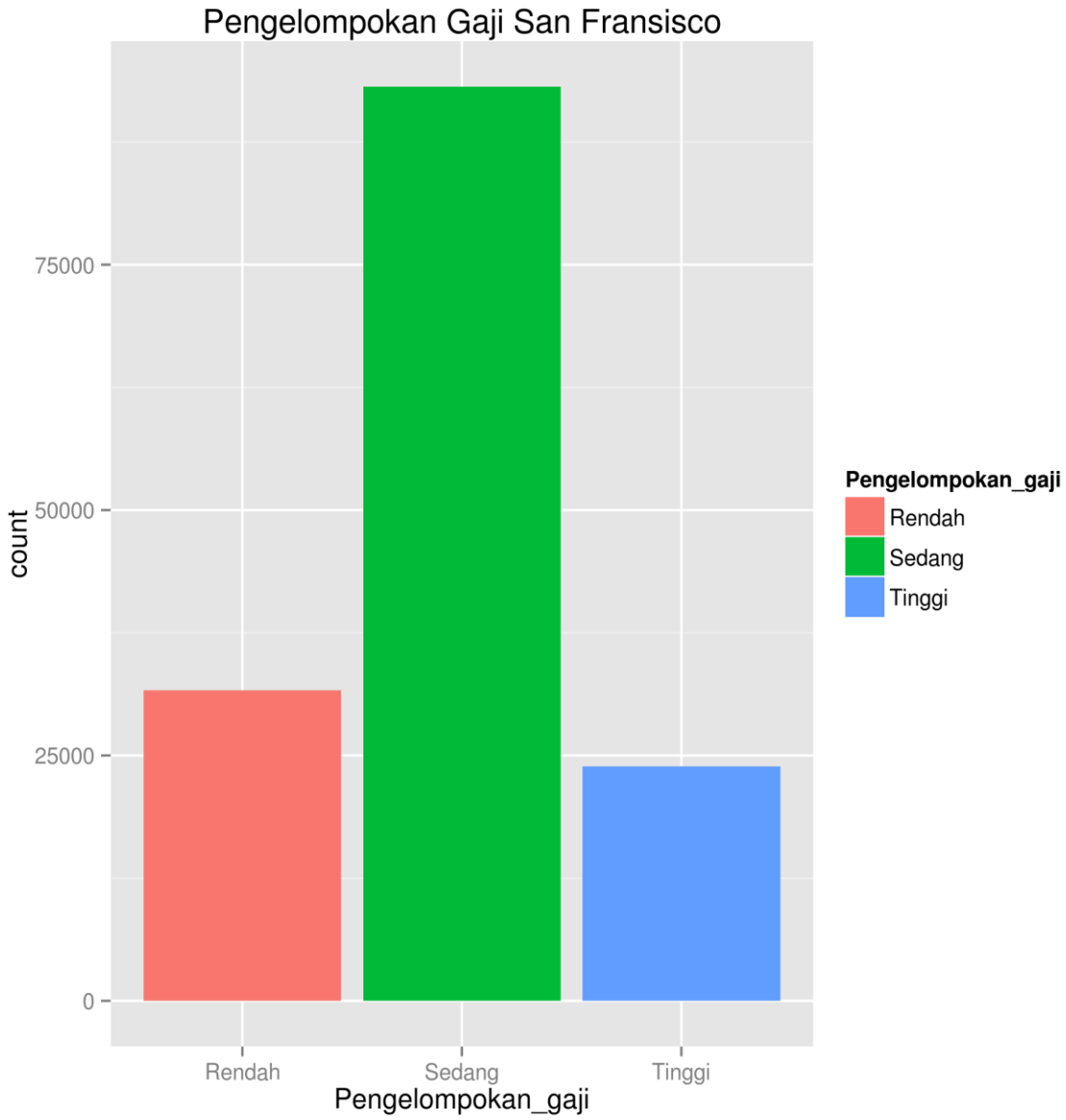
Gambar 1 Rata-Rata Gaji Penduduk San Francisco 2011-2014



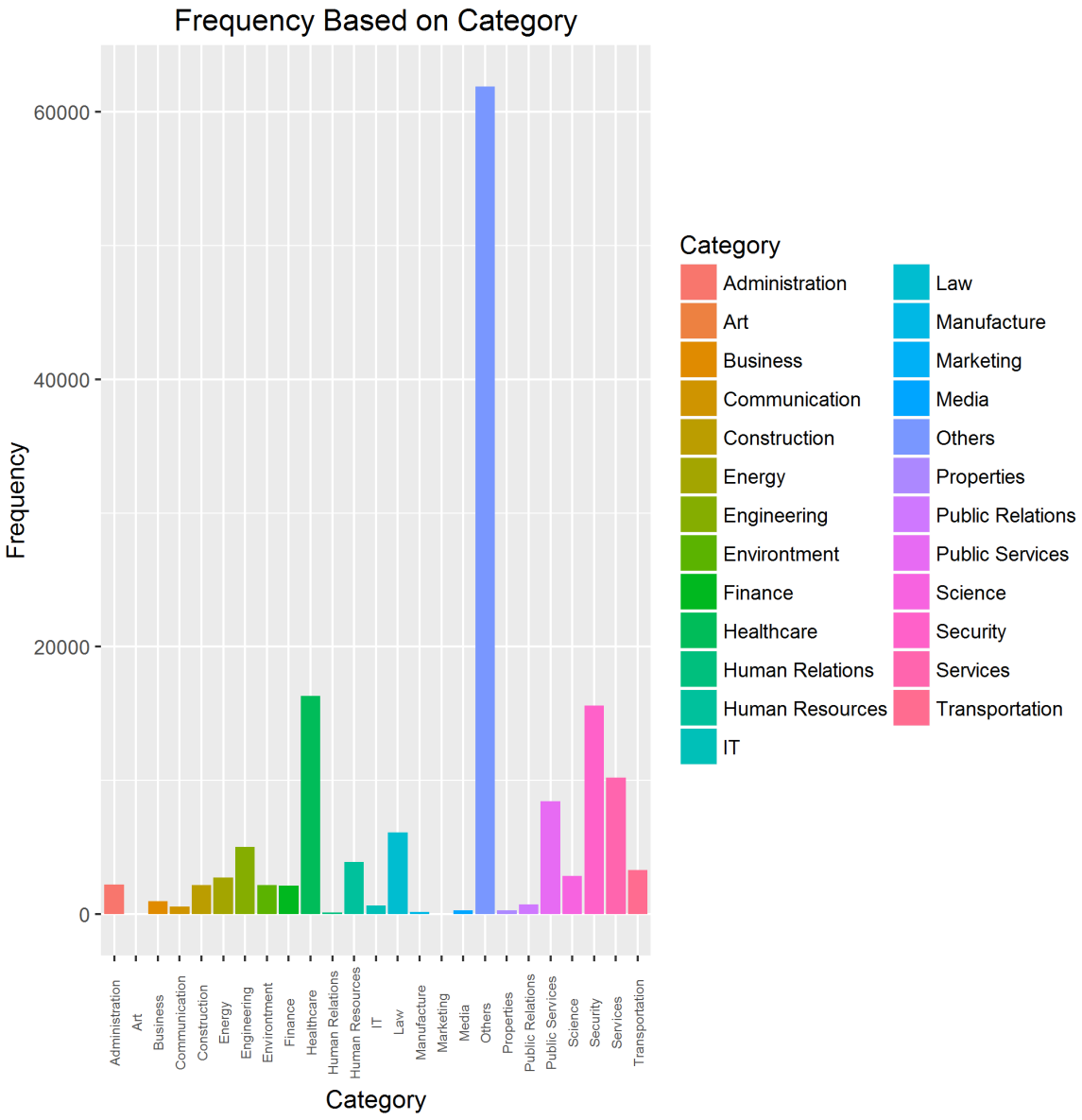
Gambar 2 Regresi Linear Gaji Penduduk terhadap Tahun



Gambar 3 Prediksi Gaji Rata-Rata Tahun 2015 dan 2016



Gambar 4 Pengelompokan Penduduk berdasarkan Gaji



Gambar 5 Pengelompokan Penduduk berdasarkan Jenis Pekerjaan