

# Salaries of San Francisco

*Mr. Garcia, Mr. Kalyan*

*January 22, 2016*

## Contents

Abstract . . . . .	1
Intro . . . . .	1
Dataset . . . . .	1
Discriminant Analysis . . . . .	3
Principal Component Analysis . . . . .	4
Correspondence Analysis . . . . .	4
Factorial Discriminant Analysis . . . . .	5

---

## Abstract

The objective of this project is to perform a descriptive analysis of the San Francisco's salaries dataset in order to find if there exists links between the variables of this dataset and answer different questions. R Studio was utilized for exploring and cleaning the data, and in order to further analyse it, SPAD software was used.

## Intro

In order for San Francisco to become a more transparent city, they decided to release a sample dataset of the Salaries of the city from the years 2010-2014. We chose this [dataset](#) in order to answer the following questions:

1. Is there any links or correlation between the Salaries of the residents of San Francisco?
2. Is there a link between Job Title and the working Status (Full-Time, Part-Time)?
3. Can we use Salary to explain the Job Title or the working Status?

In order to pursue this study, the methods "PCA, CA, and FDA" were applied for discriminant analysis.

## Dataset

### Loading

First of all the dataset was loaded, and its dimensions were explored. The dataset contains 13 variables and 148654 observations. Which we know in advance that it is more than the amount which can be read by SPAD. For this reason the data will be further explore in order to determine what are the important variables for the study.

## Description

The dataset has the following variables:

Name	Class	Description
Id	Int	Id of the dataframe
EmployeeName	Factor	Name of the employee
JobTitle	Factor	Title of the job position
BasePay	Factor	Base anual salary
OvertimePay	Factor	Total Overtime payment
OtherPay	Factor	Other payments
Benefits	Factor	Anual extra benefits received
TotalPay	Numerical	Total salariy without Benefits
TotalPayBenefits	Numerical	Total Payment including Benefits
Year	Int	Year (2010-2014)
Notes	Logical	Notes are Empty
Agency	Factor	Place (San Francisco for all the observations)
Status	Factor	Status of the Employee (Full Time, Part Time)

An exploration of the *structure* and a *summary* of the data was made in order to have insights about the data. After this we have enough information to decide which variables have to be filtered, reduced or even removed.

## Preprocessing the dataset:

1. Remove variables: `Id`, `EmployeeName`, `Notes` , `Agency` will not be useful for the analysis, so they will be removed.
2. Filter the year: The dataset contains 3 years of data, for the purpose of this study, the focus will rely on year 2014.
3. Filter variable “TotalPayBenefits”: the values below zero are going to be considered as a mistake and they will be removed.
4. Remove values: The empty values for “Status” will be removed
5. Classes: `BasePay`, `Overtime Pay`, `Other Pay` and `Benefits` will be turned into numericals.

At this point, the dataframe was reduced to: 9 variables, out of which 7 are numerical and 2 categorical, and 19000 observations.

The variable `JobTitle` (categorical), is hard to analyse by having 2159 levels, or modalities, and since SPAD's student license can not handle huge datasets, a decision was made to filter the dataset by the top 10 most popular `Job Titles` in San Francisco.

Finally the dataset ends up with the following structure and summary:

The dataframe was reduced to: 9 variables and 5282 observations.

## Data Structure

```
## 'data.frame':   5282 obs. of  9 variables:
## $ JobTitle      : Factor w/ 2159 levels "Account Clerk",...: 2038 2038 2038 2038 2038 2038 2038 2038 2038 2038 ...
## $ BasePay       : num  83997 84651 83101 82222 81838 ...
## $ OvertimePay   : num  59452 57451 52025 49521 47731 ...
```

```
## $ OtherPay      : num  68826 69563 76318 73639 74802 ...
## $ Benefits      : num  72037 73204 72029 70263 70017 ...
## $ TotalPay      : num  155765 151286 138509 131865 128545 ...
## $ TotalPayBenefits: num  194247 190147 176989 169765 166361 ...
## $ Year          : int   2014 2014 2014 2014 2014 2014 2014 2014 2014 2014 ...
## $ Status        : Factor w/ 3 levels "", "FT", "PT": 2 2 2 2 2 2 2 2 2 2 ...
```

## Data Summary

```
##                JobTitle      BasePay      OvertimePay
## Transit Operator      :1266   Min.    :   14   Min.    :   3
## Special Nurse         : 753   1st Qu.: 17184  1st Qu.:   3
## Registered Nurse      : 589   Median : 47800  Median :15834
## Public Svc Aide-Public Works: 449   Mean    : 48420  Mean    :21567
## Firefighter           : 424   3rd Qu.: 74658  3rd Qu.:40316
## Custodian             : 401   Max.    :109713  Max.    :66127
## (Other)               :1400
##      OtherPay      Benefits      TotalPay      TotalPayBenefits
## Min.    :   7   Min.    :   7   Min.    :   0   Min.    :   0
## 1st Qu.:14924  1st Qu.:22134  1st Qu.: 14859  1st Qu.: 16835
## Median :37138  Median :59250  Median : 64044  Median : 88554
## Mean    :37888  Mean    :51692  Mean    : 69099  Mean    : 90208
## 3rd Qu.:59660  3rd Qu.:78564  3rd Qu.:114865  3rd Qu.:148326
## Max.    :84216  Max.    :98639  Max.    :287480  Max.    :325718
##
##      Year      Status
## Min.    :2014      :   0
## 1st Qu.:2014      FT:2378
## Median :2014      PT:2904
## Mean    :2014
## 3rd Qu.:2014
## Max.    :2014
##
```

Now this is the dataset which will be uploaded to SPAD in order to be analyzed.

## Discriminant Analysis

After importing the data to SPAD, the first step is to generate statistics and explore our values. (SHOULD WE PRINT THE STATISTICS FOR CONTINUOUS AND CATEGORICAL VARIABLES?) The continuous variables seem to be fine, it is only important to be aware for further analysis that **TotalPay** and **TotalBenefitsPay** are the linear combination of the other continuous variables.

Then, it is important to evaluate the behaviour of the categorical variables by creating a cross table, which will be later needed for a Correspondance Analysis.

Count/weight % in row % in column	Custodian	Deputy Sheriff	Firefighter	Patient Care Assistant	Police Officer 3	Public Svc Aide- Public Works	Recreation Leader	Registered Nurse	Special Nurse	Transit Operator	Overall
	276	279	389	119	331	9	0	176	1	798	2378
<b>FT</b>	11.6	11.7	16.4	5.0	13.9	0.4	0.0	7.4	0.0	33.6	100.0
	68.8	89.7	91.7	36.3	82.8	2.0	0.0	29.9	0.1	63.0	45.0
<b>PT</b>	4.3	3.2	3.5	2.09	6.9	4.40	3.61	4.13	7.52	4.68	2.904
	31.2	1.1	1.2	7.2	2.4	15.2	12.4	14.2	25.9	16.1	100.0
	401	311	424	328	400	449	361	589	753	1266	5282
<b>Overall</b>	7.6	5.9	8.0	6.2	7.6	8.5	6.8	11.2	14.3	24.0	100.0
	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Because Special Nurse and Recreation Leader have a frequency lower than five for one of the modalities of Status, then *Yates correction* is applied. Special Nurse and Registered Nurse will now become the new category Nurse, while, Recreational Leader will be discarded, providing with these changes another variable called New\_JobTitle.

## Principal Component Analysis

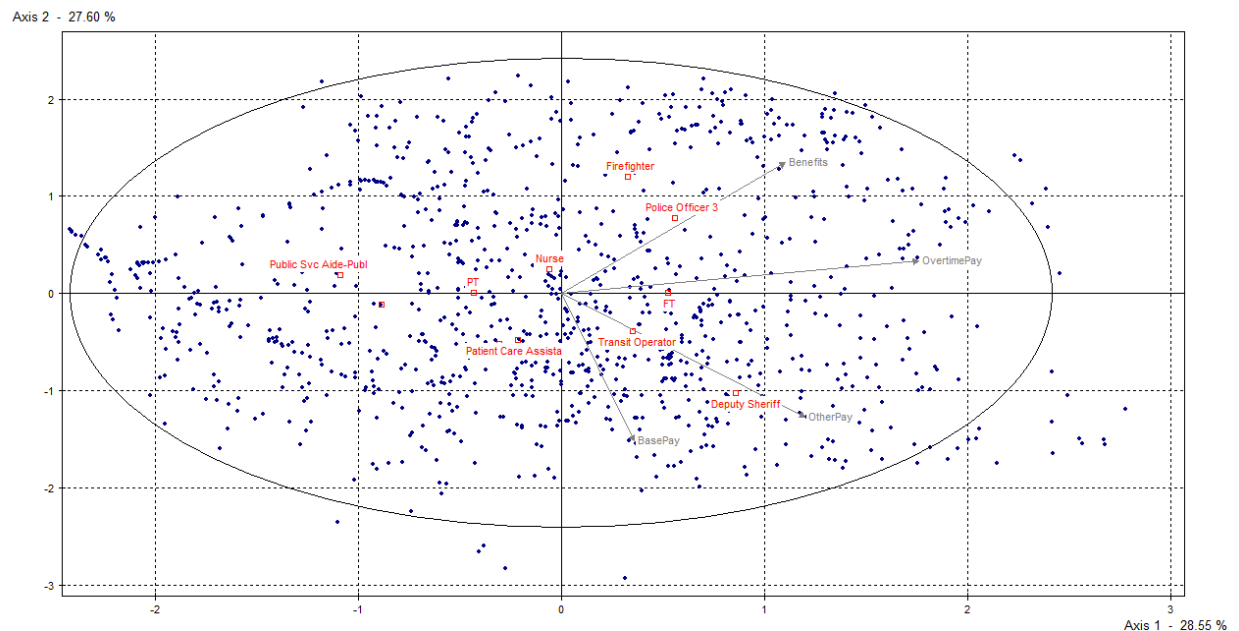
### Is there any links or correlation between the Salaries of the residents of San Francisco?

The four continuous variables selected for PCA were BasePay, OvertimePay, OtherPay, Benefits. New\_JobTitle and Status were chosen as supplementary categorical variables. The other two continuous variables TotalPay and TotalPayBenefits were removed from the analysis as they were linear combinations of the variables selected for PCA.

### Analysis

The first two principle axis capture 56% of the variation in data. Significant active variables on the first factorial plane are OvertimePay and OtherPay and significant active variables on the second factorial plane are the following variables BasePay, Benefits and OtherPay.

There is a link between OvertimePay and OtherPay on the first factorial plane where as Benefits and BasePay are in opposition on the second factorial plane. We see that Transit Operator and Deputy Sheriff have similar profile just as Firefighter and Police Officer have similar profiles, and they both tend to be FullTime, different from Public Service Aid, which tends to be PartTime.



Hereby is the [Table of Contributions](#).

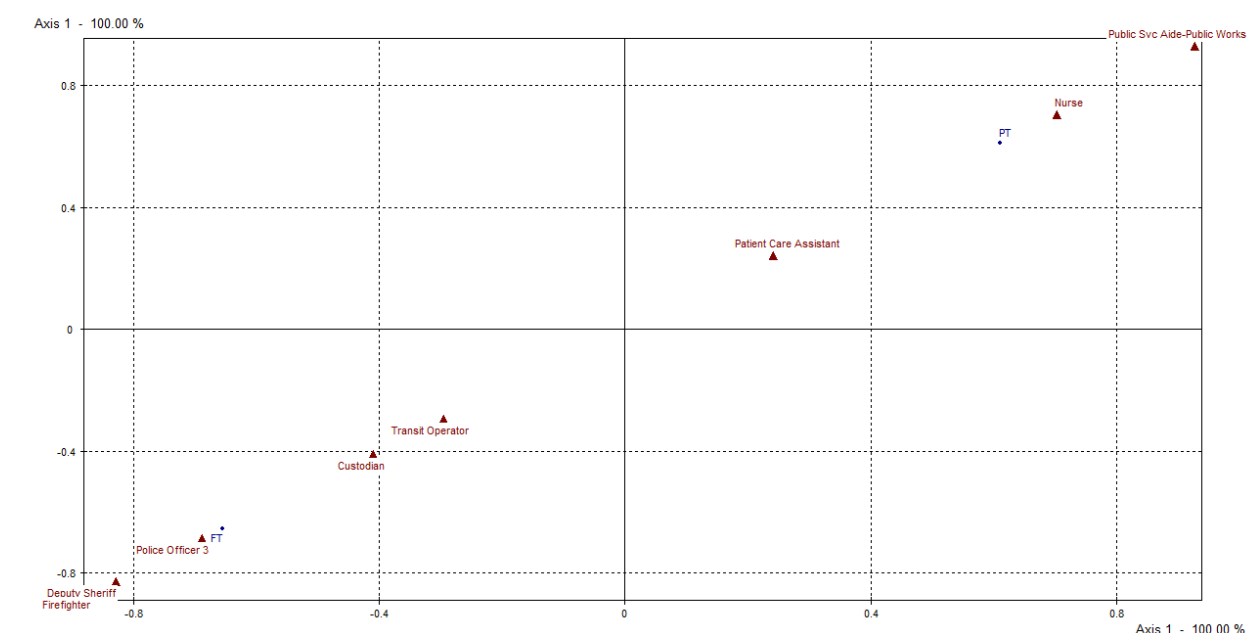
## Correspondence Analysis

### Is there a link between Job Title and the working Status (Full-Time, Part-Time)?

In order to pursue a Correspondence Analysis (CA), the first step is to revise whether it exists dependence between the two categorical variables. The variables selected for the CA are JobTitle and Status. A [Cross Table](#) was built and its Chi-square value is equal to 1969 and P-value < .001. These previous numbers

confirm a strong dependence between the variables **Job Title** and **Status**, hence a it is possible to do a Correspondence Analysis in order to answer the question.

All the variation in our data is well captured by the first principle component since one of our variables has two modalities. For the variable **JobTitle** Nurse, Public Service Aide and Fire Fighter have high contribution on the first factorial plane compared to other modalities. There exist a strong link between Full time jobs and Deputy Sheriff, Fire Fighter and Police Officer. In contrast there is also a strong link between Part time jobs and Public Service Aide and Nurse.



Hereby is the [Table of Contributions](#).

## Factorial Discriminant Analysis

### Can we use Salary to explain the Job Title or the working Status?

Given the fact that in order to do a Factorial Discriminant Analysis (FDA), is necessary to have an explain categorical variable and several explanatory continuous variables, two FDAs will be executed, one with the variable to explain being, **JobTitle** and the other being, **Status**.

### Factorial Discriminant Analysis on JobTitle

Normally a Factorial Discriminant Analysis would be capable of analysing an explain categorical variable with multiple modalities. Since the SPAD student version used for this study can not support more than two, the modalities for **JobTitle** had to be reduced. The new variable is called **Job Category** and its modalities are the following:

- **Defense:** By merging Custodian, Deputy Sheriff, Firefighters, Police Officer 3, and Transit Operator.
- **Healthcare:** Patient Care Assistant, Public Svc Aide-Public Works, Recreation Leader, Registered Nurse, and 'Special Nurse'.

The explanatory variables are **BasePay**, **OvertimePay**, **OtherPay**, **Benefits** and to explain the variable **JobCategory**. Modalities of the variable **JobCategory** are **Defense** and **Healthcare**. The model obtained is

significant as p value is less than 5%. All the explanatory variables are significant as their absolute ratios  $\geq 1.96$ . 68% of the data points are well classified by the model. Overtime pay has the highest function of Fisher.

Hereby is the [Classification Table](#).

### **Factorial Discriminant Analysis on Status**

For the Discriminant Analysis on Status in SPAD, is possible to directly use

The explanatory variables are **BasePay**, **OvertimePay**, **OtherPay**, **Benefits** and to explain the variable **Status**. Modalities of the variable **Status** are Full Time and Part Time. The model obtained is significant as p value is less than 5%. All the explanatory variables are significant as their absolute ratios  $\geq 1.96$ . 70% of the data points are well classified by the model. Overtime pay has the highest function of Fisher.

Hereby is the [Classification Table](#).