SSB DAY-5

Task -1

Use the given data as shown and compute:

Average: bonus and increment received by each employee

Draw the chart as shown in the third picture below

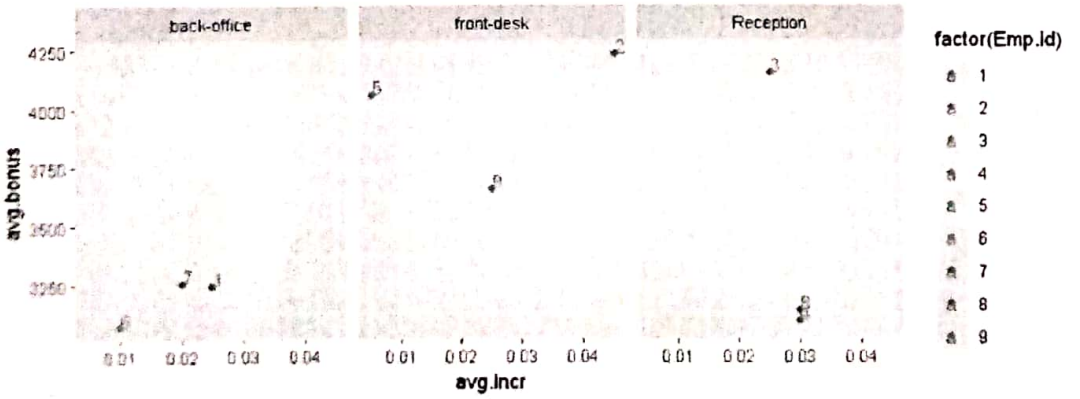
Data:

- data generation script provided

In [3]:

```
set.seed(10)
df = data.frame(Emp.id = 1:9,
                dep = sample(rep(c('front-desk', 'back-office', 'Reception'), each = 3)),
                Bonus.Cur = sample(2000:5000, replace=TRUE, size=9),
                Bonus.Prev = sample(2000:5000, replace=TRUE, size=9),
                Increment.Cur = round(runif(9, 0.0, 0.05), digits=2),
                Increment.Prev = round(runif(9, 0.0, 0.05), digits=2))
```

Emp.id	dep	Bonus.Cur	Bonus.Prev	Increment.Cur	Increment.Prev
1	back-office	3289	3196	0.01	0.04
2	front-desk	3955	4509	0.04	0.05
3	Reception	3703	4595	0.02	0.03
4	Reception	2340	3846	0.03	0.03
5	front-desk	3788	4326	0.00	0.01
6	back-office	3074	3067	0.01	0.01
7	back-office	3286	3217	0.04	0.00
8	Reception	2155	4120	0.02	0.04
9	front-desk	2792	4515	0.04	0.01



Task -2

You have ChickWeight dataset in-built in R. Use the dataset and for every diet create line plots of time variation of weights of individual chicken

Save the charts to local drive (respective) using ggsave function

Data:

`ChickWeight`

Task -3

You have been provided with a set of data on singapore tourism numbers (2010-2019). The task is to model the data for tourist inflow from across the world with respect to other independent variable(s). Formulate a strategy of how the modeling is to be done (one or multiple models??). Compare and contrast between the various models looking a the goodness of fit statistic.

Create a 3-4 slide presentation to discuss your findings with class. sum up data by ~~month~~ year

Data:

`tourism.zip`

Task -4 : Data Preparation & Modeling

`All flights out of 3 DC airports (WAS) into 3 NYC airports which were not cancelled in January 2004`

META-DATA

CRS_DEP_TIME: Departure Time
Carrier
Origin
DEST : Destination
Distance
FL_DATE: Flight Date
FL_NUM: Flight Number
Weather: 1 means weather-related delay, 0 otherwise
Day_Week: Day of the Week
Day_Of_month: Day of the Month
Tail Number
Flight.Status: binary variable (Target)

Carrier Code	Carrier Name
AA	American Airlines, Inc.
CO	Continental Air Lines, Inc.
DH	Atlantic Coast Airlines
DL	Delta Air Lines, Inc.
EV	Atlantic Southeast Airlines
FL	Airtran Airways Corporation
MQ	American Eagle Airlines,inc
OH	Comair, Inc.
RU	Continental Express Airline
UA	United Air Lines, Inc.
US	US Airways, Inc.

1. Use some functions from tidyverse package to explore and prepare data
2. Model the data using binary logistic regression
3. Predict whether a flight is likely to be delayed

Data:

Delay.csv

Task -5 : Data Preparation & Modeling

```
- Read the diabetes dataset from UCLA repository
- Use the meanings given below to name the columns
    # col-1. Number of times pregnant
    # col-2. Plasma glucose concentration a 2 hours in an
             oral glucose tolerance test
    # col-3. Diastolic blood pressure (mm Hg)
    # col-4. Triceps skin fold thickness (mm)
    # col-5. 2-Hour serum insulin (mu U/ml)
    # col-6. Body mass index (weight in kg/(height in m)^2)
    # col-7. Diabetes pedigree function
    # col-8. Age (years)
    # col-9. Diabetes (binary 1 or 0; Target variable
```

1. Use some functions from tidyverse package to explore and prepare data
2. develop a simple logistic regression to predict whether a person with given measures is likely to have diabetes

Data:

```
(script given)

 pima = read.csv("pima-indians-diabetes.csv",
            col.names=c("Pregnant","Plasma_Glucose","Dias_BP",
                        "Triceps_Skin","Serum_Insulin","BMI",
                        "DPF","Age","Diabetes"))

  head(pima)
```

Task -6 : Data Modeling

Case: Owners of a nationwide restaurant chain wants to use a data driven strategy to pick the location of their next branch. The team has been provided with a dataset with meta-data as explained above. Use this to develop a model that can help the restaurant owners to pick their next location.

You can test your model on the test data (**test.csv**)

Data:

```
studenmunds_restaurants.csv
```

```
Meta-data
```

1. Sales: gross sales volume at each chain

2. Competition: number of direct competitors in 2 mile radius
3. Population: number of people in the 3 mile radius
4. Income: avg household income

DataSource: Using Econometrics A Practical Guide A.H. Studenmund Sixth Edition; Pearson Publications

Submission: Create a one-page report summarizing your study explaining to the business what are the important factors in deciding the location and what additional data can be useful to refine your work.

Task -7

Case: Data on faculty salary of a certain university is provided. Perform a study to answer the following:

1. Is there significant difference in the salaries of males and females ?
2. Is there significant difference between the salaries of Assitant/Associate/Profs ?
3. What are the significant predictors of a faculty salary as per given data ?

Data:

`uni.csv`

DataSource: salaries data from car package in R

Submission: Create a one-page report/slide summarizing your study.