Case Study 1
New Employee's Performance

Download dataset: http://tinyurl.com/dateset-of-casestudy1
Alternative link – Click Here-Data
Note: Before you start Read complete document carefully

Problem Domain

A company hires many individuals from various colleges across the country. After a year, they do a comprehensive performance appraisal for all these individuals. The appraisal is based on their on-job performance: a combination of objective metrics met by them and subjective rating by their managers. Each candidate is rated as a best performer (BP), mid performer (MP) or a low performer (LP). During the hiring process, the company had taken these candidates through 15 different tests and an incredible set of profiling. Yet, they landed up with mid and low performers!

Data Sets

There are 666 observations with candidates' ID, Biographical information, Test scores and their performance. Identification of each candidate is coded for privacy. All biographic information i.e. name, gender, year of birth, state, degree of study, specialization of study, year of completion of college are coded as alphabets. Performance is BP, MP and LP which stand for Best performance, Medium Performance and Low performance after a year in company.
Missing Data: you will find MD in many places in all the data that stand for 'Missing Data'. It happens when some observation are lost or could not be collected or have doubt about values.

Your Job for today: **"Just draw your conclusions"**

You have to apply everything you have learnt so far for drawing your conclusions. You don't have to make any model to predict anything, you are supposed to draw the conclusions only to show the pattern and behavior of data.

Your final report should contain

1. Doc file of conclusions with required graphs, figures and statistical numbers to support your each conclusion. **Note: Your each conclusion should have supportive graphs, numbers and figures.**
2. xls sheet containing data with added sheets of your containing graphs, charts table, pivot table linked to data to reproduce graphs included in doc file
3. Script file of SAS and SQL queries to generate results and reports which you have included in doc file with conclusion.
4. Any report generated by SAS or SQL or xls (If Any)

Your conclusion can be like this
(These are just random statements for example, not the exact analysis of data, these statement might be wrong as per data)

Obsolete Conclusions

- 'State A' has candidates with good English
- Gender B of state C are Low in domain skills.
- Specialization C are good in performance.
- State D's Candidate having Average performance

Relative or comparison

- Gender A of State D are better than Gender A of State E
- Average 10th class percentage of candidates of Best performance category state wise and then gender wise.

Correlation analysis

- If English 1 of candidate is good, than his/her English 2 is also good.
- If Analytical skill of any candidate is not good, his/her domain knowledge is also not good
- In Specialization C English knowledge does not affect the performance of candidates
- In specialization X candidate performance very much depends on domain skill 2
- Candidates having overall English knowledge with average marks and average domain knowledge greater than x percentage always perform best in Companies

There can be many more such statements and conclusions, we recommend you to draw conclusions which **'surprise'** everyone and which is **meaningful.** Do not conclude something which is quite obvious or even if it is obvious, add some figures with it to make it more sensible for example

Conclusion 1 : Candidate have low analytical skill, Low domain skill and Low English score are always perform low in organization.

Conclusion 2: Candidate have analytical skill performance lower than 40%, average domain skill lower than 45% and average English score lower than 50% always perform low in organization.

Conclusion 2 make more sense than conclusion 1. So whenever you make any conclusion, instead of saying low, good or poor, try to say it with numbers.

Privacy act of data

You will always get data on public domain with encoded patterns, there will be no identification of any individual for privacy purpose. You are never allowed to use same data for any commercial purpose. Such data is always used for Academic purpose. You cannot claim of anything about any 'Individual' (name) or any particular (State, gender, degree of year)
However when you will be working with any real time scenario or project for clientele or in organization as "Data Analyst" you will have no coding until client permit.

You may map data to any random identifications to have a feeling of real data or project (See the mapping in **Tips sections**)

Tips

1. **Revise your Statistics:**
   Revise your statistics, Mean, mode, Median, variance, standard deviation, range, interquartile range, correlation coefficient, skewness, bar graph, pie chart, box plot, histogram. Revise what each of these terms says and how you can use them for analysis and conclusions

2. **Handling Missing Data:**
   There are two ways to handle Missing data
   a. Delete the COMPLETE observation where data is missing. We apply this when we have enough data
   b. Replace missing data with some 'aggregate value'. This needs again lots of other efforts to find the similar cases to aggregate values. This is done when we have very little amount of data.

   "Never ever replace missing data with 0, or any random values". This will adds Outliers in you data and will misleads your conclusions
   For this situation, we recommend you to delete all the observations which have **'MD'**

3. **Replace coded data:** If it helps you to understand data and have real time feeling you may replace coded data with following mapping
   Mapping
   States

Code	State	Code	State
A	Uttar Pradesh	O	Assam
B	Maharashtra	P	Punjab
C	Bihar	Q	Chhattisgarh
D	West Bengal	R	Haryana
E	Madhya Pradesh	S	Jammu and Kashmir
F	Tamil Nadu	T	Uttarakhand
G	Rajasthan	U	Himachal Pradesh
H	Karnataka	V	Tripura
I	Gujarat	W	Meghalaya
J	Andhra Pradesh	X	Manipur[β]
K	Odisha	Y	Nagaland
L	Telangana	Z	Goa
M	Kerala	AB	Arunachal Pradesh
N	Jharkhand	AA	Mizoram

Year: Y1 =1981, Y2=1982 ….. Y9 =1989
Y10 =2000, Y11 =2001
Y16 =2006 … Y20 = 2010

Gender: A=MALE, B=Female

Mapping of Degree

W : Science
X : Engineering
Y : Finance
Z : Diploma

Mapping of Specialization

A : Biotech
B : CSE
C : Civil
D : Arch
E : Chemical
F : Math
G : ECE
H : Marin
I : Accounting
J : IT
K : ELE
L : ME