**Project guide**

**Data exploration(**<https://blog.csdn.net/EnjoySmile/article/details/50611690>

1. **)**
2. Choose and load the R dataset corresponding to your group subject’s and Identify which variables your data set are numeric, and which are categorical (factors) if applicable

You can use the command data() to list all the datatsets,

You can create a table by using the data(name of your data set). to get more information on you data set tape its name in the help window.

You can use the command mode()

1. Generate summary level descriptive statistics: Show the mean, median, 25th and 75th quartiles, min, and max for each of the applicable variables in your data set
2. Determine the frequency for each of one of the categorical variables.
3. Determine the frequency for each of the one of the categorical variables, by a different categorical variable.

You can use the command levels to discover the different modalities of the categorical variables. Use also Summary() or table() for question 4.

1. **Graphic data representation**
   1. **Univariate study**
2. Use the commands pie() barplot and dotchart() to represent the categorical data. Comment
3. Create a graph for each single numeric variable. (histogram)
4. Create a graph for each single numeric corresponding to different categorical variable (histogram)
   1. **Bivariate study**
5. Use the command plot or sunflowerplot to plot the scatterplot of the dependent and independant variables. What is the diffrence between these two commands ? Comment your results.
   1. **Graphic representation for the different data categories**
6. Represent the scatter plot for the the dependent and independant variables for each data category. Comment
7. **Regression Analysis**
   1. **Testing hypothesis**
8. Use the Khi-deux test to verify the indenpendant of each data category.
9. Testing the Standard Assumptions of linear regression
   1. **Build the model the regression model**
   2. **Verify model significance ( Model validation)**

**Datatsets**

|  |  |  |
| --- | --- | --- |
| **dataset** | **Dataset number** | **group** |
| **airquality** | **1** |  |
| **quakes** | **2** |  |
| **ChickWeight** | **3** |  |
| **iris** | **4** |  |
| **longley** | **5** |  |
| **esoph** | **6** |  |
| **occupationalStatus** | **7** |  |
| **mtcars** | **8** |  |

PPT（3/12） presentation(17/12)

R program（problem、aim、）

**#occupationalStatus**

**##I**

**#查看数据集的大小和结构，维度、名称、属性。**

**dim(occupationalStatus)**

**names(occupationalStatus)**

**str(occupationalStatus)**

**attributes(occupationalStatus)**

**#查看数据集的前几行和后几行**

**occupationalStatus[1:5,]**

**head(occupationalStatus)**

**tail(occupationalStatus)**

**#用summary()看数据的分布情况**

**summary(occupationalStatus)**

**#看平均值、中位数、极差，四分位数和百分位数**

**mean(occupationalStatus)**

**median(occupationalStatus)**

**range(occupationalStatus)**

**quantile(occupationalStatus)**

**min(occupationalStatus)**

**max(occupationalStatus)**

**#查看方差，计算密度估计值(密度函数线)**

**var(occupationalStatus)**

**density.default(occupationalStatus)**

**#茎叶图**

**stem(occupationalStatus)**

**##II**

**#饼图（对于因子，可以先算频数，然后画饼图或条形图）**

**table(occupationalStatus)**

**pie(table(occupationalStatus))**

**#条形图**

**barplot(table(occupationalStatus))**

**#克利夫兰点图**

**dotchart(table(occupationalStatus))**

**#绘制直方图**

**hist(occupationalStatus)**

**#马赛克图**

**mosaicplot(occupationalStatus)**

**#??What is the diffrence between these two commands ? Comment your results.**

**#plot图 （就是马赛克图）**

**plot(occupationalStatus)**

**#sunflowerplot() 函数**

**sunflowerplot(occupationalStatus)**

**#III回归模型的评估**

**#拟合统一的关联模型，将对角线效果分开 Fit a uniform association model separating diagonal effects**

**#残差与拟合（https://blog.csdn.net/qq\_35837578/article/details/88357551）**

**#（Residuals vs Fitted）普通残差与拟合值的残差图：这是残差与真实值之间的关系画图**

**#（Normal Q-Q plot）正态QQ的残差图：用来检测其残差是否是正态分布的**

**#（Scale-Location plot）标准化残差的开方与拟合值的残差图：这个图是用来检查等方差假设的**

**#（Residuals vs Leverage）残差与杠杆：这种图的意义在于检查数据分析项目中是否有特别极端的点**

**Diag <- as.factor(diag(1:8))**

**Rscore <- scale(as.numeric(row(occupationalStatus)), scale = FALSE)**

**Cscore <- scale(as.numeric(col(occupationalStatus)), scale = FALSE)**

**modUnif <- glm(Freq ~ origin + destination + Diag + Rscore:Cscore,**

**family = poisson, data = occupationalStatus)**

**summary(modUnif)**

**plot(modUnif)**

**# 4 plots, with warning about h\_ii ~= 1**