STAT3622 Quiz 2
(Open Book, But No Group Discussion)
Due on April 18 Midnight 12am

1. Use PCA to visualize the potential clusters with a high-dimensional dataset.

   (a) Load the dataset `covtype_pca.csv` into `R`, perform the *K*-means clustering with the number of clusters set as 3, and save the output cluster labels.

   (b) Conduct PCA on the data, obtain the first two principal components (PC1 and PC2), and define a data frame with three columns as PC1, PC2 and the corresponding cluster labels.

   (c) Based on the predefined data frame, use `ggplot` to visualize the clusters with respect to the first two principal components as follows.

2. Peripheral arterial disease (PAD) is a common cardiovascular disease which affects about 10% of the general population worldwide. In recent years, the newly developed drug-coated balloons (DCBs) and drug-eluting stents (DESs) with paclitaxel in the femoropopliteal arteries have shown substantial improvements in clinical efficacy compared with standard percutaneous transluminal angioplasty (PTA). However, the safety of long-term use of paclitaxel DCB and DES has raised great concerns.

   The CSV file 'jaha_paclitaxel.csv' includes the number of all-cause deaths and overall number of patients at 1 year, 2 years and 4 or 5 years in the paclitaxel intervention and control arms for 28 randomized controlled trials (RCTs).

   - Study: trial name;
   - P.Events: number of all-cause death in the paclitaxel group;
   - P.Total: overall number of patients in the paclitaxel group;
   - C.Events: number of all-cause death in the control group;
   - C.Total: overall number of patients in the control group;
   - Period: follow-up period.

   (a) Conduct both fixed and random effects meta-analyses on the **2-year** all-cause mortality. Use relative risk (RR) as the summary measure and Mantel-Haenszel method for weighting. Output the pooled treatment effect estimates in both fixed and random effects models.

   (b) Draw the forest plot.

   (c) Draw a funnel plot and report the Egger's test. Do you think publication bias is a problem in this meta-analysis?

Hint on K-means clustering

```
## k-means algorithm
set.seed(2021)
cl = kmeans(df,3)

## PCA
df.cov = cov(df)
df.eigen = eigen(df.cov)

## PVE plot and cumulative PVE plot
PVE = df.eigen$values/sum(df.eigen$values)
```

```
PVEplot = qplot(1:length(PVE),PVE)+
  geom_line()+xlab('Principal Component')+
  ylab('PVE')
PVEplot

cumPVEplot = qplot(1:length(PVE),cumsum(PVE))+
  geom_line()+xlab('Principal Component')+
  ylab('cumPVE')
cumPVEplot
```