1.

a. The probablilty of the event of woman who resigned belongs to store C can be represented by P{C|W}

Using Bayes theorem

P{C|W}=P{C,W}/P{W}

where P{W} is probability of woman and P{C,W} is probability of woman belonging to C.

P{C,W}=P{W|C}*P{C}

P{C}=100/50+75+100

P{W|C}=70/100

P{W}=0.5*50+0.6*75+0.7*100/50+75+100

P{C|W}=(100/225)*(70/100)/(0.5*50+0.6*75+0.7*100/50+75+100)

=70/140 = **0.5**

b.

	POSITIVE	NEGATIVE
TEST POSTIVE	TRUE POSITIVE:0.95	FALSE POSITIVE:0.01
TEST NEGATIVE	FALSE NEGATIVE:0.05	TRUE NEGATIVE:0.99

probability that a person has the disease given that the test is positive can be represented by P{P|TP} where P{P} is probability of having disease and P{TP} is probability of getting tested positive.

Using Bayes' Theorem

P{P|TP}= P{P,TP}/P{TP}

Probablity of a person having disease=0.5/100=0.005

P{P,TP}= P{TP|P}*P{P}=0.95*0.005

P{TP}=P{TP|P}*P{P}+P{TP|N}P{N}=0.95*0.005+0.01*0.995

P{P|TP}=0.95*0.005/0.95*0.005+0.01*0.995

=0.00475/0.0147= **0.323**

c. Probability that Atlanta Braves wins the division is represented by P{AB}

P{AB}=P{Atlanta Braves wins all three games}+P{Atlanta Braves wins two games and wins playoff games with SF giants or LA dodgers who won all the three games}+ P{Atlanta Braves wins two games and SF giants or LA dodgers win two of the three games}+P{ Atlanta Braves

wins one game and wins playoff games with SF giants or LA dodgers who won two of the three games }

P{AB}= $\frac{1}{2^3} + 3 * \frac{1}{2^3} * \frac{1}{2} * 2 * \frac{1}{2^3} + 3 * \frac{1}{2^3} * \frac{1}{2} * 3 * 2 * \frac{1}{2^3} + 3 * \frac{1}{2^3} * 2 * 3 * \frac{1}{2^3} = \frac{\mathbf{76}}{\mathbf{128}}$ **= 0.59375**

d. Probability of additional playoff game P{A}= probability that Atlanta Braves wins one of the three games and either SF giants or LA Dodgers win two of the three games+ probability that Atlanta Braves wins two of the three games and either SF giants or LA Dodgers win three games

P{A}= P(Atlanta Braves wins one of the three games)*P(SF giants or LA Dodgers win both the remaining games)+P(Atlanta Braves wins two of the three games)*P(SF giants or LA Dodgers win three games)

P{A}=3*($\frac{1}{2}$)($\frac{1}{2}$)($\frac{1}{2}$)*2*3*($\frac{1}{2}$)($\frac{1}{2}$)($\frac{1}{2}$)+3*($\frac{1}{2}$)($\frac{1}{2}$)($\frac{1}{2}$)*2*($\frac{1}{2}$)($\frac{1}{2}$)($\frac{1}{2}$)= $\frac{18}{64} + \frac{6}{64} = \frac{\mathbf{3}}{\mathbf{8}}$ **= 0.375**

2.

a. Maximum likelihood estimation of Poisson distribution:

P(x_i=k)= $\frac{\lambda^k e^{-\lambda}}{k!}$

Probability Density Function $f(x_1, x_2, .., x_n|\lambda)$ is given by

$f(x_1, x_2, .., x_n|\lambda)$ = $\frac{\lambda^{x_1} e^{-\lambda}}{x_1!} \frac{\lambda^{x_2} e^{-\lambda}}{x_2!}$ …. $\frac{\lambda^{x_n} e^{-\lambda}}{x_n!}$

= $\frac{e^{-n\lambda} \lambda^{\Sigma x_i}}{\Pi(x_i!)}$

$$lnf = -n\lambda + ln\lambda(\mathbf{\Sigma} x_i) - ln\Pi(x_i!)$$

Differentiating with respect to λ and equating it to 0

$$\frac{dlnf}{d\lambda} = -n + \Sigma x_i/\lambda = 0$$

So, maximum likelihood estimator of lambda

$$\hat{\lambda}_{\boldsymbol{MLE}} = \frac{\boldsymbol{\Sigma x_i}}{\boldsymbol{n}}$$

b. Maximum likelihood of θ_j for multinomial distribution:

Probability Density function $f(x_1, x_2, \dots, x_n; n, \theta_1, \theta_2, \dots, \theta_n) = \frac{n!}{x_1! x_2! \dots x_n!} \Pi \theta_j^{x_j}$

$$lnf = \ln(n!) - \Sigma ln(x_j)! + \Sigma ln\theta_j^{x_j}$$
$$= \ln(n!) - \Sigma ln(x_j)! + \Sigma x_j ln\theta_j$$

As a property of multinomial distribution $\boldsymbol{\Sigma}\theta_i = 1$ Therefore using Lagrange multipliers to maximize the equation

$$l(\theta_1, \theta_2, \dots, \theta_n, \lambda) = l(\theta_1, \theta_2, \dots, \theta_n) + \lambda(1 - \Sigma\theta_i)$$

Applying langrange multiplier

$$lnf = \ln(n!) - \Sigma ln(x_j)! + \Sigma x_j \, ln\theta_j + \lambda(1 - \Sigma\theta_i)$$

By differentiating with respect to θ_j and equating the equation to 0, we get $\theta_j = \frac{(x_j)}{\lambda}$

$$\boldsymbol{\Sigma}\theta_i = 1 \; So, \boldsymbol{\Sigma}\frac{(x_j)}{\lambda} = 1; \; \lambda = \boldsymbol{\Sigma}(x_j) = n$$

Maximum likelihood estimation of $\boldsymbol{\theta_j = \theta_{MLE} = \frac{x_j}{n}}$

c. Maximum likelihood estimation of μ and σ^2 for Gaussian distribution:

$$N(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\Pi}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

Probability density function $f(x_1, x_2, .., x_n|\mu, \sigma^2) =$
$\frac{1}{\sigma\sqrt{2\Pi}} \exp\left(-\frac{(x_1-\mu)^2}{2\sigma^2}\right) \frac{1}{\sigma\sqrt{2\Pi}} \exp\left(-\frac{(x_2-\mu)^2}{2\sigma^2}\right) .. \frac{1}{\sigma\sqrt{2\Pi}} \exp(-\frac{(x_n-\mu)^2}{2\sigma^2})$

$$= \frac{1}{(\sigma\sqrt{2\Pi})^n} \exp\left(-\Sigma\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

$$lnf = -\Sigma\frac{(x_i-\mu)^2}{2\sigma^2} - nln(\sigma\sqrt{2\Pi})^1$$

Differentiating *lnf* function with respect to μ and equating it to zero:

$$\frac{dlnf}{d\lambda} = \Sigma(-2)\frac{x_i-\mu}{2\sigma^2} = 0$$

$$\Sigma(x_i - \mu) = 0$$

$$n\mu = \Sigma x_i$$

$$\boldsymbol{\mu_{MLE} = \frac{\Sigma x_i}{n}}$$

Differentiating *lnf* function with respect to σ and equating it to zero:

$$\frac{dlnf}{d\,\sigma} = -\frac{n}{\sigma} + \Sigma\frac{(x_i-\mu)^2}{\sigma^3} = 0$$

$$\Sigma\frac{(x_i-\mu)^2}{\sigma^3} = \frac{n}{\sigma}$$

$$\sigma^2 = \Sigma\frac{(x_i-\mu)^2}{n}$$

$$\boldsymbol{\sigma_{MLE} = \sqrt{\Sigma\frac{(x_i-\mu)^2}{n}}}$$

3. Minimizing reconstruction error in Principal Component Analysis:

$$J = \frac{1}{N}\Sigma_{n=1:N}||x^n - \tilde{x}^n||^2$$

$$J = \frac{1}{N}\Sigma_{n=1:N}\left(x^n - \tilde{x}^n\right)(x^n - \tilde{x}^n)^T$$

Where $\tilde{x}^n = \Sigma_{i=1:M} z_i^n u_i + \Sigma_{i=M+1:D} b_i u_i \; and \; x^n = \Sigma_{i=1:D}(x^{nT}u_i)u_i$

$$J = \frac{1}{N}\Sigma_{n=1:N}(\Sigma_{i=1:D}(x^{nT}u_i)u_i - \Sigma_{i=1:M} z_i^n u_i + \Sigma_{i=M+1:D} b_i u_i)\,(\Sigma_{i=1:D}(x^{nT}u_i)u_i - \Sigma_{i=1:M} z_i^n u_i + \Sigma_{i=M+1:D} b_i u_i)^T$$

$$J = \frac{1}{N}\Sigma_{n=1:N}(\Sigma_{i=1:M}(x^{nT}u_i - z_i^n)u_i - \Sigma_{i=M+1:D}(x^{nT}u_i - b_i)u_i)(\Sigma_{i=1:M}(x^{nT}u_i - z_i^n)u_i - \Sigma_{i=M+1:D}(x^{nT}u_i - b_i)u_i)^T$$

$$J = \frac{1}{N}\Sigma_{n=1:N}(\Sigma_{i=1:M}(x^{nT}u_i - z_i^n)(x^{nT}u_i - z_i^n)^T - \Sigma_{i=M+1:D}(x^{nT}u_i - b_i)(x^{nT}u_i - b_i)^T$$

a. Minimizing the reconstruction error with respect to z_j^n.

$$\frac{dJ}{dz_j^n} = \frac{-2}{N}\left(x^{nT}u_j - z_j^n\right)$$

$$\frac{dJ}{dz_j^n} = 0$$

$$which\ implies\ \boldsymbol{z_j^n = x^{nT}u_j}$$

b. Minimizing the reconstruction error with respect to b_j

$$\frac{dJ}{db_j} = \frac{d\,\frac{-1}{N}\Sigma_{n=1:N}\Sigma_{i=M+1:D}(x^{nT}u_i - b_i)(x^{nT}u_i - b_i)^T}{db_j}$$

$$= \frac{2}{N}\Sigma_{n=1:N}\Sigma_{i=M+1:D}\,(x^{nT}u_i - b_i) = 0$$

$$\Sigma_{n=1:N} x^{nT}u_i = b_i(N)$$

$$\boldsymbol{b_i = \frac{\Sigma_{n=1:N}x^{nT}u_i}{(N)} = \overline{x^{nT}}u_i}$$

c. Optimal $\tilde{x}^n$:

Substituting $z_j^n = x^{nT}u_j$ and $b_i = \overline{x^{nT}}u_i$ in the equation $\tilde{x}^n = \Sigma_{i=1:M} z_i^n u_i + \Sigma_{i=M+1:D} b_i u_i$, we get

Optimal $\tilde{x}^n = \Sigma_{i=1:M}(x^{nT}u_i)u_i + \Sigma_{i=M+1:D}\overline{(x^{nT}}u_i)u_i$

Optimal $x^n - \tilde{x}^n = \Sigma_{i=1:D}(x^{nT}u_i)u_i - \Sigma_{i=1:M}(x^{nT}u_i)u_i - \Sigma_{i=M+1:D}\overline{(x^{nT}}u_i)u_i$

$$= \Sigma_{i=M+1:D}(x^{nT}u_i)u_i - \Sigma_{i=M+1:D}\overline{(x^{nT}}u_i)u_i$$

$$= \Sigma_{i=M+1:D}(x^{nT}u_i - \overline{x^{nT}}u_i)u_i$$

$$= \Sigma_{i=M+1:D}((x^n - \overline{x^n})^T u_i)u_i$$

d.

From the above derivations , the cost function can be written as

$$J = \frac{1}{N}\Sigma_{n=1:N}||x^n - \tilde{x}^n||^2 = \frac{1}{N}\Sigma_{n=1:N}\Sigma_{i=M+1:D}||((x^n - \overline{x^n})^T u_i)u_i||^2$$

$$=\frac{1}{N}\Sigma_{n=1:N}\Sigma_{i=M+1:D}((x^n - \overline{x^n})^T u_i)u_i)^T\left((x^n - \overline{x^n})^T u_i)u_i\right)$$

$$=\frac{1}{N}\Sigma_{n=1:N}\Sigma_{i=M+1:D}(u_i)^T((x^n - \overline{x^n})^T u_i)^T\left((x^n - \overline{x^n})^T u_i)(u_i\right)$$

$$=\frac{1}{N}\Sigma_{n=1:N}\Sigma_{i=M+1:D}(u_i)^T\left((x^n - \overline{x^n})^T u_i)((x^n - \overline{x^n})^T u_i)^T(u_i\right)$$

$$=\frac{1}{N}\Sigma_{n=1:N}\Sigma_{i=M+1:D}(u_i)^T(x^n - \overline{x^n})\left((x^n - \overline{x^n})^T(u_i)\right)$$

$$=\frac{1}{N}\Sigma_{n=1:N}\Sigma_{i=M+1:D}(u_i)^T(x^n - \overline{x^n})(x^n - \overline{x^n})^T(u_i)$$

Taking covariance matrix $S = \frac{1}{N}\Sigma_{n=1:N}(x^n - \overline{x^n})(x^n - \overline{x^n})^T$

$$J = \Sigma_{i=M+1:D}(u_i)^T S(u_i)$$

The above cost function is to be minimized for u_j where u_j is orthonormal and is subject to $u_j(u_j)^T$=1. Using langrange multiplier λ_j

$$J = (u_j)^T S(u_j) + \lambda_j(1 - (u_j)^T u_j)$$

Derivating the reconstruction error with respect to u_j

$$\frac{dJ}{du_j} = (u_j)^T S - \lambda_j(u_j)^T = 0$$

$$(u_j)^T S = \lambda_j(u_j)^T$$

By taking transpose on both the sides,

$$Su_j = \lambda_j u_j$$ (S is a symmetry matrix)

So u_j is the eigen vector of covariancve matrix S. The larger the value of eigen values, the more covariance is retained in the reduced matrix. So, the minimum value of *J* can be achieved by electing eigenvectors corresponding to the *D* - M smallest eigenvalues, and the eigenvectors defining the principal subspace are those corresponding to the M largest eigenvalues.

4.

a. $J = \Sigma_{n=1:N}\Sigma_{k=1:K} r^{nk}||x_n - \mu^K||^2$

Minimizing the cost function with respect to μ^K

$$r^{nk} = 1\ if\ x^n\ belongs\ to\ K\ cluster; 0\ otherwise$$

$$\frac{dJ}{d\mu^K} = \frac{d\Sigma_{n=1:N} r^{nk}||x_n - \mu^K||^2}{d\mu^K}$$

$$\frac{dJ}{d\mu^K} = -2\Sigma_{n=1:N} r^{nk}(x_n - \mu^K)$$

Equating the above statement to zero

$$2\Sigma_{n=1:N} r^{nk}(x_n - \mu^K) = 0$$

$$\mu^K \Sigma_{n=1:N} r^{nk} = \Sigma_{n=1:N} r^{nk} x_n$$

$$\boldsymbol{\mu^K = \frac{\Sigma_{n=1:N} r^{nk} x_n}{\Sigma_{n=1:N} r^{nk}}}$$

b. Proof that K-means algorithms converge to local minima:

$$\frac{dJ}{d\mu^K} = -2\Sigma_{n=1:N} r^{nk}(x_n - \mu^K)$$

Taking the differential of the above equation

$$\frac{d^2 J}{d^2 \mu^K} = 2$$

The second order derivative of the function with respect to μ^K is a positive value. This proves that the function converges to a local minima in finite steps.

In every iteration of K-means the centers are moving closer the clusters and the points with are closest to a particular center are grouped together. So k-means bound to converge to local minima in finite steps

c. I believe, Average Linkage will give clusters similar to K-Means clustering

In K-means, the points with minimal distance between the centres are grouped together into a cluster. And the centre of a cluster is calculated by mean of the points present in the cluster. The main goal in clustering is to minimize the cost function which is given by

$$J = \Sigma_{n=1:N} \Sigma_{k=1:K} r^{nk} ||x_n - \mu^K||^2$$

And $\mu^K = \frac{\Sigma_{n=1:N} r^{nk} x_n}{\Sigma_{n=1:N} r^{nk}}$

And by average linkage hierarchial clustering, clusters with minimum average distance between all the points of the clusters are grouped together.

i.e, $\frac{1}{mp} \boldsymbol{\Sigma\Sigma} ||\boldsymbol{x_i - y_j}||$

Ultimately, in both the methods the average of all the input points present are considered to form clusters.

d. Single linkage will separate two moons efficiently.

Single linkage forms clusters in the form of long chain of points when the points are placed within close proximities. Here every points within each moon is relatively very close to adjacent points with in the same moon compared to the points in the other moon. So single linkage will form long chain clusters by grouping points locally within each moon in finite steps.

5.

1. Algorithm used for K-medoids:

   Initialize cluster centers to random K points from the dataset

While centers calculated for the current iteration is not same as the previous iteration's:

The data points belong to that cluster where the distance metric is minimum

Previous centres= centres

Compute centres for all the clusters by taking mean of all points in cluster

if any one of the clusters is empty

run the k-medoids for K-1 number of clusters

else

choose cluster center as the data point in the cluster which is close to computed center

I am choosing complete linkage as the distance metric to form the clusters. I have tried hamming distance, Manhattan distance and Euclidean distance. Euclidean and Manhattan distances give similar results to K-Means.

2.

K-MEDOIDS:

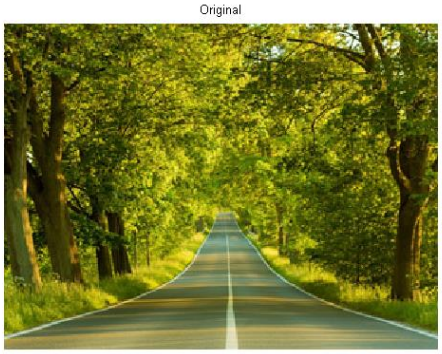
3.

For K=3: (Execution time is 8.12 seconds)

Original

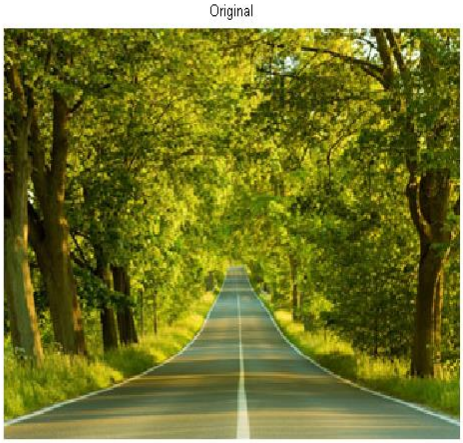
K-medoids

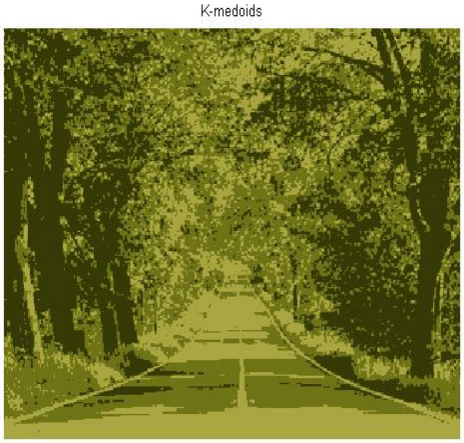
For K=16: (Execution time is 84.3512 seconds)



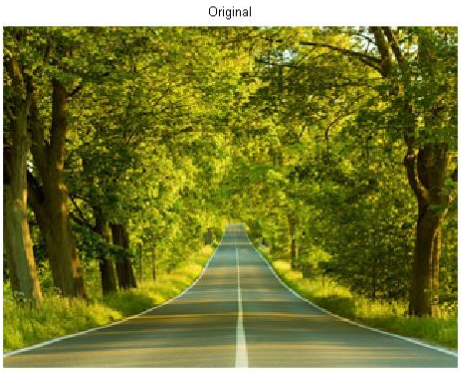
With increase in the number of clusters, the image reconstructed looks similar to the original image i.e., reconstruction error is reducing. But the execution time also increases with increase in the number of clusters

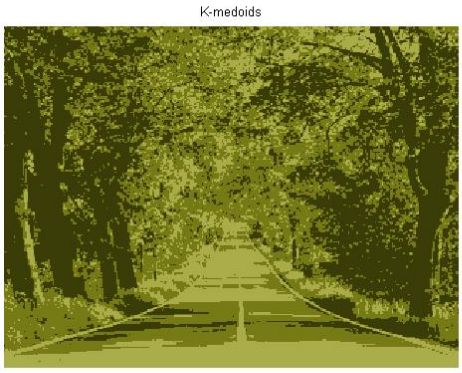
4. K-medoids by taking first K points as cluster centers:
   Time taken to converge is 18.0784





K-medoids by taking dividing the dataset into five sets and taking the first point in each subset as cluster centers: Time taken to converge is 5.027 seconds

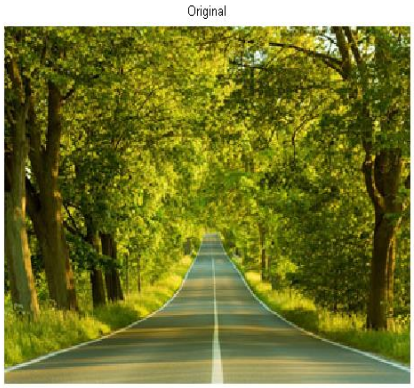


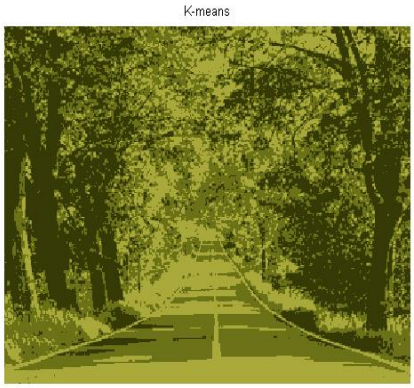


K-MEANS:

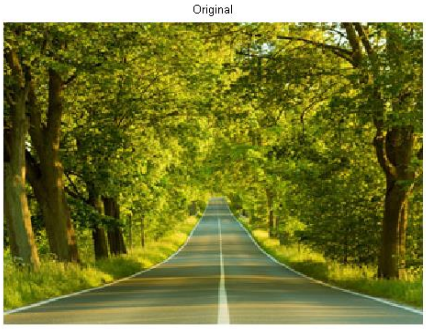
3.

For K=3: (Execution time is 1.5757 seconds)

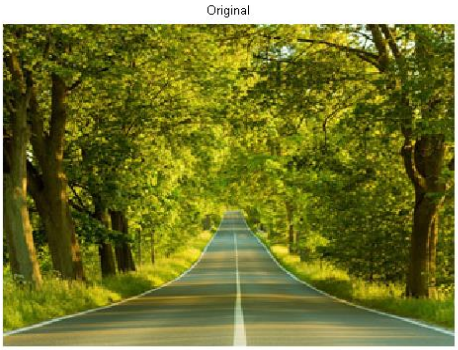


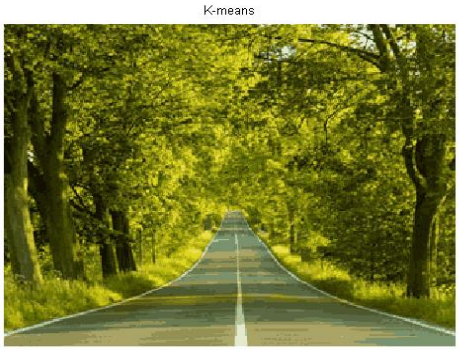


For K=16: (Execution time is 5.2795 seconds)



For K=32: (Execution time is 20.529 seconds)





Similar to K-medoids, with increase in the number of clusters, the image reconstructed looks similar to the original image and the execution time increases with increase in the number of clusters