ST451 Course Project

Classification, Clustering and Linear Regression using Bayesian Methods

Table of Contents

1 **PART 1**

- i) Aim ……………………………………………………………………………….Pg. 3
- ii) Introduction …………………………………………………………………….Pg. 3
- iii) Data
  - (1) Exploratory Data Analysis……………………………………………………….Pg. 4
- iv) Background and Methodology
  - (1) Data Pre-processing……………………………………………………………Pg. 4
  - (2) Bayesian Logistic Regression………………………………………………….Pg. 4
  - (3) Cluster Analysis
    - (a) Gaussian Mixture Model…………………………………………………….Pg. 5
    - (b) Bayesian Gaussian Mixture Model…………………………………………..Pg. 6
    - (c) K Means Clustering………………………………………………………….Pg. 6
- v) Results
  - (1) Bayesian Logistic Regression…………………………………………………Pg. 7
  - (2) Cluster Analysis
    - (a) Gaussian Mixture Model……………………………………………………Pg. 8
    - (b) Bayesian Gaussian Mixture Model…………………………………………..Pg.10
    - (c) K Means Clustering…………………………………………………………Pg.10
- vi) Conclusion…………………………………………………………………….Pg.11

2 **PART 2**

- i) Aim……………………………………………………………………………….Pg. 11
- ii) Introduction………………………………………………………………........Pg.11
- iii) Data
  - (1) Exploratory Data Analysis…………………………………………………..Pg.11
- iv) Methodology
  - (1) Multiple Linear Regression…………………………………………………Pg.12
  - (2) Ridge Regression……………………………………………………………Pg.12
  - (3) Bayesian Linear Regression………………………………………………....Pg.12
- v) Results
  - (1) Ridge Regression……………………………………………………………Pg.13
  - (2) Linear Regression……………………………………………………………Pg.13
  - (3) Bayesian Linear Regression…………………………………………………Pg.13
- vi) Conclusion……………………………………………………………………Pg.14

3 References…………………………………………………………………………Pg.15

4 Appendix………………………………………………………………………... Pg.16

PART 1

Aim

1. To predict whether a person earns more or less than $50k given different factors
2. To understand the most important determinants of income
3. To determine how income classes vary with age and number of working hours

Introduction

Rising income inequalities are a major concern for many countries as it is leading to unstable economies, higher crime rates, lower social mobility and poor public health. Thus, one of the United Nations Sustainable Development goal is to reduce income inequalities by 2030. This project uses a 1994 US census survey to identify the important determinants of income and the different clusters of income based on age and number of working hours per week. This analysis will determine the driving factors behind difference in income levels and thereby help government and development organisations formulate policies to achieve more equitable incomes.

The UCI 'Adult Data Set' consists of 13 attributes which help to determine whether or not a person earns more than $50k a year. The first task for this project is to fit a classification model to determine whether a person earns more or less than $50k and to understand the most important determinants of these income levels. The second task is to perform cluster analysis to understand how age and number of working hours per week affects a person's income level. We will use Bayesian methods for both these tasks because of the comparative advantages they provide. We will compare the results of Bayesian methods with classical methods to check how the results compare.

Advantages of Bayesian methods over classical methods

1. Bayesian methods help in incorporating prior knowledge or beliefs to the observed data when training models. Bayesian credible intervals allow additional information that is external to the sample to be incorporated. This additional information may improve accuracy and credibility of estimations
2. Bayesian methods allow for out of sample prediction and so they are useful when modelling is performed on a regular basis as in census surveys.
3. Small sample inference for Bayesian methods proceeds in the same manner as large sample. Thus, Bayesian methods are useful when only one sample exists and additional samples cannot be gathered.
4. Bayesian analysis also can estimate any functions of parameters directly, without using the "plug-in" method as the Maximum Likelihood estimates
5. Bayesian methods provide a convenient setting for a wide range of models, such as hierarchical models and missing data problems. Bayesian techniques like Markov Chain Monte Carlo make computations tractable for virtually all parametric models.

Data

This project used the dataset 'Adult Income' from the UCI repository. The dataset is credited to Ronny Kohavi and Barry Becker and was drawn from the 1994 United States Census Bureau data and involves using personal details such as education level to predict whether an individual will earn more or less than $50,000 per year.The dataset provides 14 input variables that are a mixture of categorical, ordinal, and numerical data types. There are a total of 32,561 rows of data. The complete list of variables is as follows:

- Age.
- Workclass
- Final Weight
- Education
- Education Number of Years
- Marital-status
- Occupation
- Relationship
- Race
- Sex
- Capital-gain
- Capital-loss
- Hours-per-week
- Native-country

Exploratory Data Analysis

- Age distribution is right skewed and not symmetric. Most people are aged between 20 and 50. The minimum and maximum ages are 17 and 90 respectively
- Most people work about 30-40 hours per week
- Over 45% of the people are married
- There are more males in the data than females. The number of females is almost half the number of males
- More than half the people are white
- People with income level< $50k have a median age of 34 while those with income level >=$50k is about 42

Please check the appendix for exploratory data visualisations

Background and Methodology

Data Pre-processing

Our data consists of 9 categorical variables which cannot directly be used for performing logistic regression. To use these in the logistic regression equation, we convert all the categorical variables into factors using the Label Encoder. We also drop any null values from the data using the dropna() function. Since there is a heavy imbalance in the number of observations for each class, we tried to use SMOTE technique, a synthetic oversampling technique to balance the classes so that the prediction accuracy for each class would be the same. However, due to the large size of the matrices, we had to use a different train-test split proportion which lead to an overall lower accuracy. Thus, we decided not to use this method for our main analysis. Please find the codes for SMOTE data in the appendix section of the notebook. After cleaning the data, we go ahead and create 70-30 train-test split. We then use the train data to fit the Bayesian Logistic Regression model.

Bayesian Logistic Regression

Logistic regression is named after the function used at the core of the method, the logistic function. The logistic function is an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1.

Logistic regression uses input values (x) are combined linearly using weights or coefficient values to predict an output value (y). A key difference from linear regression is that the output value being modelled is a binary values (0 or 1) rather than a numeric value. Below is an example logistic regression equation:

y = e^(b0 + b1*x) / (1 + e^(b0 + b1*x))

Where y is the predicted output, b0 is the intercept term and b1 is the coefficient for the single input value (x). Each column in the input data has an associated b coefficient (a constant real value) that must be learned from the training data.

We now turn to a Bayesian treatment of logistic regression. In logistic regression, we need maximize the likelihood function p(y|β0,β1,x) (find Maximum Likelihood Estimate) to find the coefficients for our model. That is, you find the weights β0,β1 that maximizes how likely your observed data is. There is no closed form solution to the MLE, so you need to use iterative methods. This gives you a single point estimate of our weights. In Bayesian logistic regression, you start with an initial belief about the distribution of p(β0,β1). Then $p(\beta_0,\beta_1|x,y) \propto p(y|\beta_0,\beta_1,x)p(\beta_0,\beta_1)$. That is, the posterior, which is our updated belief about the weights given evidence, is proportional to our prior (initial belief) times the likelihood. We can't evaluate the closed form posterior, but can approximate using Laplace approximation. In Laplace approximation, we approximate the posterior p(β|x,y) with a Gaussian.

Finally, we use the Newton-Raphson algorithm[1] to find the estimates by minimising the negative log posterior. Thus, using this method, we find the posterior mean, standard error and hence confidence interval of the model's coefficients. Once we have these values, we will identify the significant variables for our model. We will also compare the results obtained using the Bayesian method to the classical logistic regression method (which uses Maximum Likelihood to estimate coefficients).

Evaluating performance

We plot the AUC curve to evaluate the accuracy that the model achieves. AUC - ROC curve is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability. AUC tells how much the model is capable of distinguishing between classes. Higher the AUC, the better the model is at predicting 0s as 0s and 1s as 1s. By analogy, the Higher the AUC, the better the model is at distinguishing between people of income >$50k and <=$50k.

Cluster analysis

Cluster analysis is a class of techniques that are used to classify objects or cases into relative groups called clusters. In cluster analysis, there is no prior information about the group or cluster membership for any of the objects. Cluster analysis involves formulating a problem, selecting a distance measure, selecting a clustering procedure, deciding the number of clusters, interpreting the profile clusters and finally, assessing the validity of clustering. In this project, we use 3 clustering methods:

a. *Gaussian Mixture Models*

A Gaussian mixture model is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters.
The GaussianMixture object implements the expectation-maximization (EM) algorithm for fitting mixture-of-Gaussian models. The EM algorithm initiates the parameters and iteratively updates from old parameters to new parameters until the log likelihood or the parameters converge. We compute the Bayesian Information Criterion to assess the number of clusters in the data.
The GaussianMixture comes with different options to constrain the covariance of the difference classes estimated: spherical, diagonal, tied or full covariance.

- spherical: each cluster k has covariance σ_k^2I
- tied: full covariance matrix but the same across clusters
- diag: diagonal covariance matrix, different for each cluster
- full: full covariance matrix, different for each cluster

b. *Bayesian Gaussian Mixture Models*

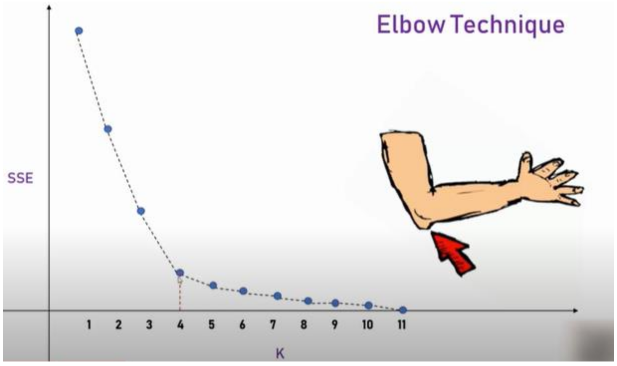
The approach so far was Bayesian with respect to the variable but not the parameter θ. For a fully Bayesian approach priors on θ should be specified. The posterior is not available in closed form. We can therefore consider a variational approximation[2]. We can apply mean field approximation and the outcome will be analogous to the EM algorithm as we repeat the procedure until convergence.

c. *K Means Clustering*

[1] The **Newton-Raphson method** is a way to quickly find a good approximation for the root of a real-valued function f(x) = 0. It uses the idea that a continuous and differentiable function can be approximated by a straight line tangent to it.

[2] Variational approximations is a body of deterministic techniques for making approximate inference for parameters in complex statistical models.

The objective of K-means is to group similar data points together and discover underlying patterns. To achieve this objective, K-means looks for a fixed number (*k*) of clusters in a dataset. A cluster refers to a collection of data points aggregated together because of certain similarities. We define a target number *k*, which refers to the number of centroids needed in the dataset. The K-means algorithm identifies *k* number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. The *'means'* in the K-means refers to averaging of the data; that is, finding the centroid. To find the optimal number of clusters we should use, we use the *'Elbow Method'*. We calculate the sum of squared errors (ie. the sum of squared distance of each data point from the centroid of each cluster) for different values of k. When we plot the sum of squared errors (SSE) against different values of k, we will observe that as k increases, the SSE decreases. This is because at some point, we can consider all of data points as individual clusters and at this point, the SSE would be 0. Thus, the general guideline is to find the 'elbow' or the point after which SSE decreases by small amounts.



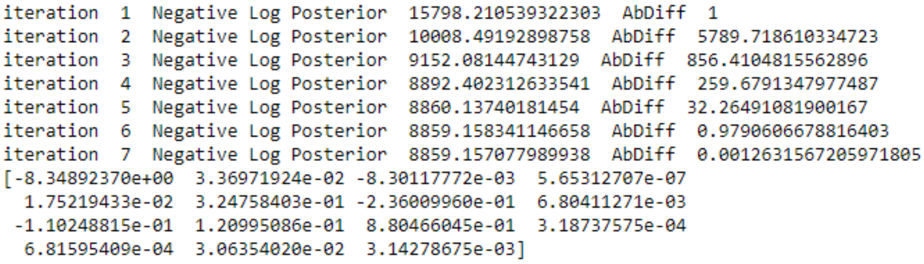
Graph 1: SSE vs k for an eleven data point dataset

As seen from Graph 1, the SSE reduces to zero as the number of clusters becomes equal to the number of data points. Thus, we choose the value for k after which the SSE decreases by small amounts. In this case, it would be k=4.

Results

Bayesian Logistic Regression

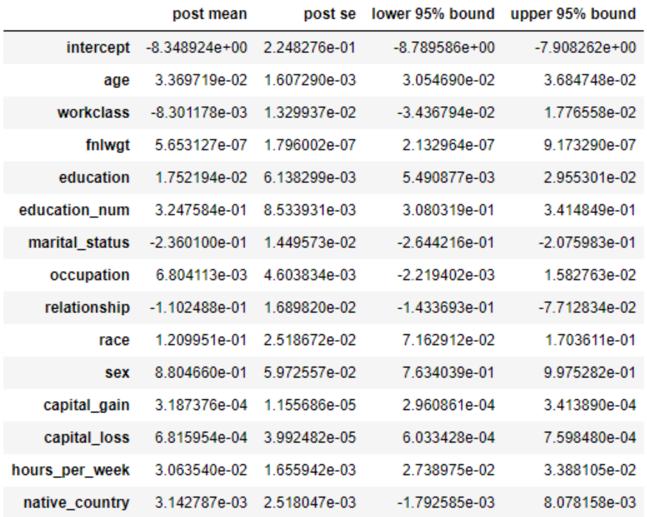
After running the Newton Raphson Bayesian Logistic Regression function, we obtain the following results after 7 iterations



```
iteration  1  Negative Log Posterior  15798.210539322303  AbDiff  1
iteration  2  Negative Log Posterior  10008.49192898758  AbDiff  5789.718610334723
iteration  3  Negative Log Posterior  9152.08144743129  AbDiff  856.4104815562896
iteration  4  Negative Log Posterior  8892.402312633541  AbDiff  259.6791347977487
iteration  5  Negative Log Posterior  8860.13740181454  AbDiff  32.26491081900167
iteration  6  Negative Log Posterior  8859.158341146658  AbDiff  0.9790606678816403
iteration  7  Negative Log Posterior  8859.157077989938  AbDiff  0.0012631567205971805
[-8.34892370e+00  3.36971924e-02 -8.30117772e-03  5.65312707e-07
  1.75219433e-02  3.24758403e-01 -2.36009960e-01  6.80411271e-03
 -1.10248815e-01  1.20995086e-01  8.80466045e-01  3.18737575e-04
  6.81595409e-04  3.06354020e-02  3.14278675e-03]
```

Image 1: Beta estimates using Newton-Raphson Bayesian Logistic Regression

The list in image 1 is the list of coefficients for the Bayesian logistic regression model that we obtained after the convergence of the absolute difference between the old negative log posterior and new negative log posterior. We also find the 95% confidence intervals for these estimates and then display the result along with the posterior mean and posterior standard error for each estimated coefficient.

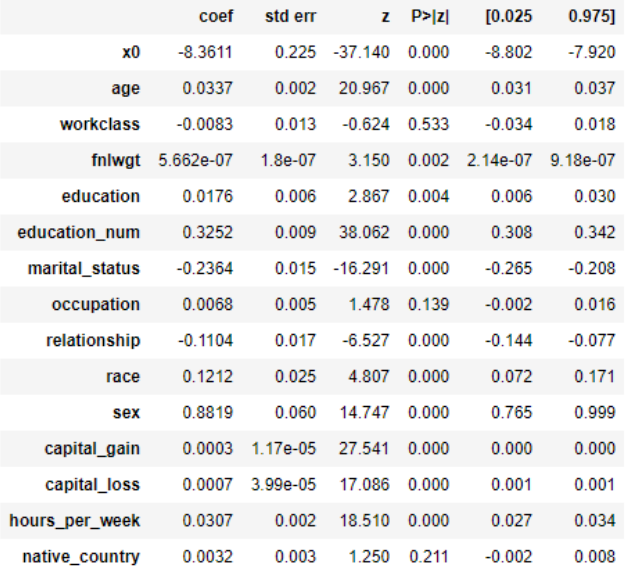


	post mean	post se	lower 95% bound	upper 95% bound
intercept	-8.348924e+00	2.248276e-01	-8.789586e+00	-7.908262e+00
age	3.369719e-02	1.607290e-03	3.054690e-02	3.684748e-02
workclass	-8.301178e-03	1.329937e-02	-3.436794e-02	1.776558e-02
fnlwgt	5.653127e-07	1.796002e-07	2.132964e-07	9.173290e-07
education	1.752194e-02	6.138299e-03	5.490877e-03	2.955301e-02
education_num	3.247584e-01	8.533931e-03	3.080319e-01	3.414849e-01
marital_status	-2.360100e-01	1.449573e-02	-2.644216e-01	-2.075983e-01
occupation	6.804113e-03	4.603834e-03	-2.219402e-03	1.582763e-02
relationship	-1.102488e-01	1.689820e-02	-1.433693e-01	-7.712834e-02
race	1.209951e-01	2.518672e-02	7.162912e-02	1.703611e-01
sex	8.804660e-01	5.972557e-02	7.634039e-01	9.975282e-01
capital_gain	3.187376e-04	1.155686e-05	2.960861e-04	3.413890e-04
capital_loss	6.815954e-04	3.992482e-05	6.033428e-04	7.598480e-04
hours_per_week	3.063540e-02	1.655942e-03	2.738975e-02	3.388105e-02
native_country	3.142787e-03	2.518047e-03	-1.792585e-03	8.078158e-03

Image 2: Posterior Mean, Posterior Standard Error and 95% confidence intervals of estimated coefficients

From the table in Image 2, it is clear that the significant variables are intercept, age, fnlwgt, education, education_num, marital_status, relationship, race, sex, capital_gain, capital_loss, hours_per_week (since 0 does not lie in their confidence intervals). Thus, age, final weight (ie. the estimated number of people each row in the data represents), education level, marital status, relationship, race, sex, capital gain and loss and working hours per week are good determinants of the income class of any individual in the dataset. Work class, Occupation and Native Country do not have a significant contribution in determining whether a person will earn more or less than $50k.

We also compare our results with the classical Logistic regression model.

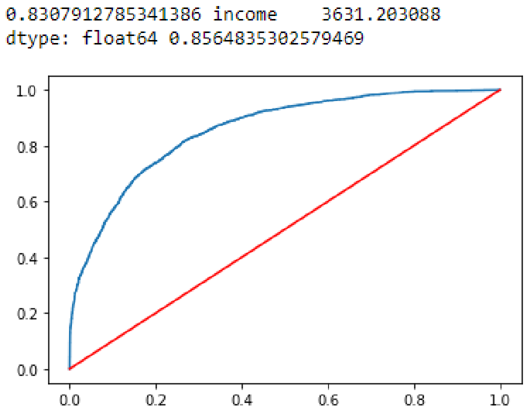


	coef	std err	z	P>|z|	[0.025	0.975]
x0	-8.3611	0.225	-37.140	0.000	-8.802	-7.920
age	0.0337	0.002	20.967	0.000	0.031	0.037
workclass	-0.0083	0.013	-0.624	0.533	-0.034	0.018
fnlwgt	5.662e-07	1.8e-07	3.150	0.002	2.14e-07	9.18e-07
education	0.0176	0.006	2.867	0.004	0.006	0.030
education_num	0.3252	0.009	38.062	0.000	0.308	0.342
marital_status	-0.2364	0.015	-16.291	0.000	-0.265	-0.208
occupation	0.0068	0.005	1.478	0.139	-0.002	0.016
relationship	-0.1104	0.017	-6.527	0.000	-0.144	-0.077
race	0.1212	0.025	4.807	0.000	0.072	0.171
sex	0.8819	0.060	14.747	0.000	0.765	0.999
capital_gain	0.0003	1.17e-05	27.541	0.000	0.000	0.000
capital_loss	0.0007	3.99e-05	17.086	0.000	0.001	0.001
hours_per_week	0.0307	0.002	18.510	0.000	0.027	0.034
native_country	0.0032	0.003	1.250	0.211	-0.002	0.008

Image 3: Results from classical logistic regression

From Image 3 we can see that the results we obtain from the classical logistic regression model are very similar to that obtained from the Bayesian Logistic regression model. This model also suggests that work class, occupation and native country do not have a significant contribution in determining whether a person will earn more or less than $50k.

Evaluating Performance



```
0.8307912785341386 income    3631.203088
dtype: float64 0.8564835302579469
```

Image 4: AUC curve for Bayesian Logistic Regression

Image 4 shows that the accuracy of the Bayesian Logistic Regression model is 83.079%. This means that 83.079% of the data points are classified correctly.

Cluster Analysis

Gaussian Mixture Models

When we fit the Gaussian Mixture Models using the variables ‘age’ and ‘number of working hours’, we obtain the following mean and covariances for each variable in each cluster.

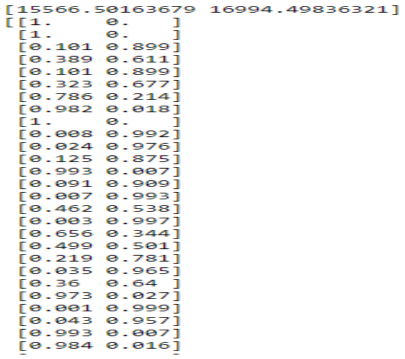
Mean	**Age**	**Working Hours**
Cluster 1	48.75928675	42.80637245
Cluster 2	29.20480623	38.25493091

Table 1: GMM Means

Cluster 1	**Age**	**Working Hours**	**Cluster 2**	**Age**	**Working Hours**
Age	129.53279952	-57.4828995	**Age**	54.771982	32.53030069
Working Hours	-57.4828995	193.74152851	**Working Hours**	32.53030069	104.48203475

Table 2: GMM Cluster 1 and 2 Covariances

The GMM method does not necessarily allocate individuals to each cluster with certainty, it allocates them with some probabilities. After soft allocation we obtain:



```
[15566.50163679 16994.49836321]
[[1.        0.       ]
 [1.        0.       ]
 [0.101 0.899]
 [0.389 0.611]
 [0.101 0.899]
 [0.323 0.677]
 [0.786 0.214]
 [0.982 0.018]
 [1.        0.       ]
 [0.008 0.992]
 [0.024 0.976]
 [0.125 0.875]
 [0.993 0.007]
 [0.091 0.909]
 [0.007 0.993]
 [0.462 0.538]
 [0.003 0.997]
 [0.656 0.344]
 [0.499 0.501]
 [0.219 0.781]
 [0.035 0.965]
 [0.36  0.64 ]
 [0.973 0.027]
 [0.001 0.999]
 [0.043 0.957]
 [0.993 0.007]
 [0.984 0.016]
```

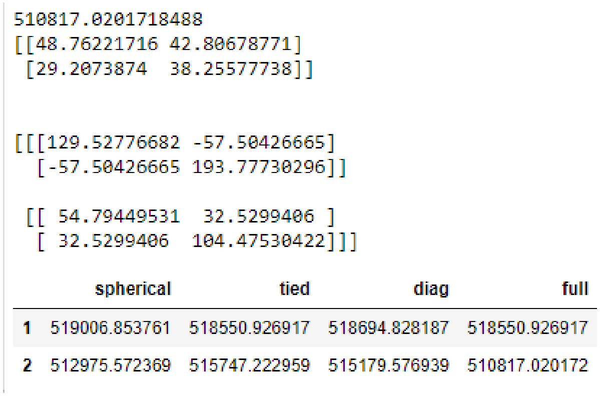
Image 5: Soft allocation of data points to clusters

This method gives the probability of each data point belonging to either cluster. Here, we are displaying the probabilities for data points between 22650 and 22677. For eg: The probability that the

22650th point belongs to cluster 1 is 1 and the probability that it belongs to cluster 2 is 0. Similarly, the probability that the 22650th point belongs to cluster 1 is 1 and belonging to cluster 2 is 0. The probability that the 22652th term belongs to cluster 1 is 0.101 while it belonging to cluster 2 is 0.899. Thus, there is a higher chance that the 22652th data point belongs to the second cluster.

Model Search

We need to fit models with different numbers of cluster and different type of covariance matrices to identify the best one. This is done via the BIC (the smaller the better in this case).



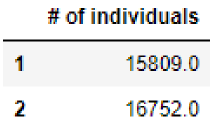
```
510817.0201718488
[[48.76221716 42.80678771]
 [29.2073874  38.25577738]]


[[[129.52776682 -57.50426665]
  [-57.50426665 193.77730296]]

 [[ 54.79449531  32.5299406 ]
  [ 32.5299406  104.47530422]]]
```

	spherical	tied	diag	full
1	519006.853761	518550.926917	518694.828187	518550.926917
2	512975.572369	515747.222959	515179.576939	510817.020172

Image 6: BIC for different number of clusters and covariances

510817.020172 is the lowest BIC out of all iterations. Looking at the table above, this value is achieved at k=2 and covariance type='full', thus we will use this to fit our Bayesian Gaussian Mixture Model. The first list is the mean of age and hours_per_week for each cluster. The second list gives us the covariances for age and hours_per_week for each cluster.

Now, we fit the Bayesian Gaussian Mixture Model with 2 clusters and covariance type ‘full’ as we obtained the lowest BIC for this combination in the previous section.



	# of individuals
1	15809.0
2	16752.0

Image 7: Number of individual in each cluster

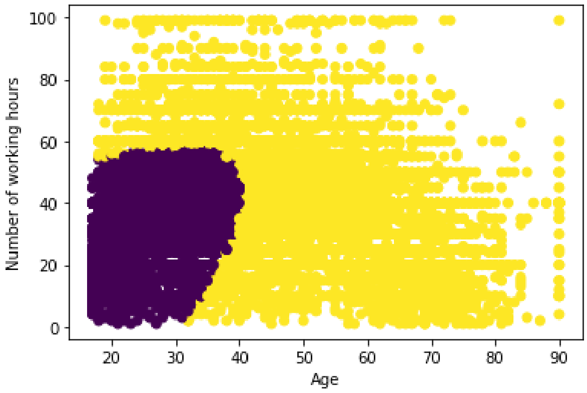


Image 8: Scatter Plot for age vs number of working hours

Plotting the fitted Bayesian Gaussian Mixture, we get the scatter plot as above. This scatter plot shows that there are two clusters of incomes. Young people (up till age 40) who work the lesser number of hours (up to 60 hours) fall into the low income class. Some individuals between ages 30 and 40 can belong to higher income groups despite working lower hours while some can belong to low income groups despite working long hours. After age 40, irrespective of the number of hours a person works, they would belong to a higher income class.

K Means Clustering

Let us also use the KMeans Clustering Algorithm to find the different clusters in the data. We will first try to find the number of clusters we should use. This is obtained using the ELBO method. In this method, we first find the sum of squared errors (SSE) for different number of clusters and then choose the k from which the SSE begins to fall. If we consider the line that we obtain on the cluster-SSE graph as an arm, then this point is the elbow of that arm.

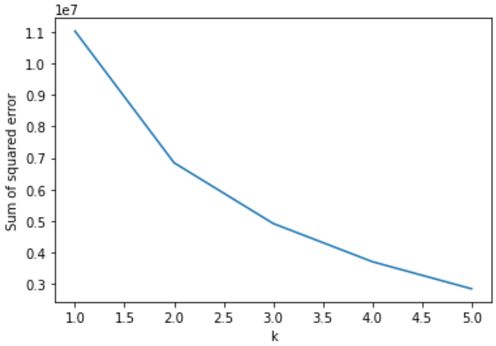


Image 9: Number of clusters vs SSE graph

Plotting the SSE for different k values shows that the 'elbow' point is at k=2. Thus, we should use 2 clusters.

After the running the K Means clustering algorithm on our data, the following scatter plot of age vs number of working hours is obtained

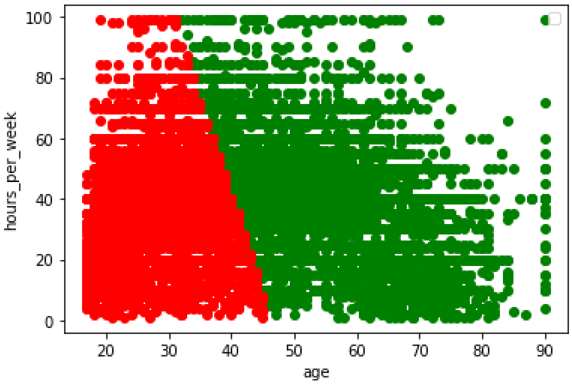


Image 10: Age vs hours_per_week scatter plot after K means clustering

This plot shows that the first cluster of incomes is associated with lower age. People of lower ages tend to belong to low income class irrespective of the number of hours they work. This, however, changes around age 35, when people who put more number of hours per week belong to higher income level. By age 45, people can expect to be in higher income levels with even lower number of working hours.

Conclusion

From our analysis, we can conclude that age, final weight (ie. the estimated number of people each row in the data represents), education level, number of years spent in education, marital status, relationship, race, sex, capital gain and loss and working hours per week are important determinants of the income class of individuals. The Bayesian logistic regression model suggests that as people spend more years in education and gain higher levels of education, the probability that they will earn more than $50k increases. This means that policy makers must focus on improving access to higher education and incentivising people to remain in education for longer. The analysis also suggests that race and gender have a significant impact on income class. White people and men have a greater probability of being in the higher income class. Thus, policy makers must research about the issue to uncover underlying discrimination and biases that are preventing women and people of colour from earning higher incomes. Another result is that married people and people with relationship ‘wife’ have the lowest chance of being in higher income groups. Therefore, government should welfare schemes for families and provide additional sources of income to married women as they stand the lowest chance of being in higher income groups. The model also suggests that older people have a higher chance of being in high income groups, thereby implying that the government needs to improve access of younger people to higher incomes through skill development and education policies.

The cluster analysis shows that there are two classes of individuals when comparing age vs number of working hours. It suggests that the first class of individuals, presumably the lower income class, belong to younger age groups and work shorter number of hours per week. As the age of individuals increase, they fall into a higher income class irrespective of the number of hours they work. Thus, government and planning authorities must improve access of younger people to higher paying jobs.

PART 2

Aim: 1. To fit a Linear and Ridge Regression model through the data and evaluate their performance 2. To fit a Bayesian Linear Regression model to predict bike-sharing sales and understand the driving factors behind the sale of bike-sharing

Introduction

As urbanisation and modernisation reach unprecedented levels, road congestion has become a modern day menace. Heavy traffic is associated with air pollution, safety risks, and losses in terms of accessibility, economic competitiveness, sustainable growth and social cohesion. If we are determined to make our cities attractive and sustainable, we must respond to these challenges. One effective way of dealing with these issues is promoting bike sharing. Bike sharing has several benefits including transport flexibility, reductions to vehicle emissions, health benefits, reduced congestion and fuel consumption, and financial savings for individuals. Some studies also report that bike share has been found to positively contribute towards additional revenues for small and medium enterprises. Therefore, in this project we will try to understand what the driving factors behind sale of bike-sharing are so that companies and policy makers can design strategies to increase the dependence on such modes of transport.

To understand the driving factors behind sales of bike-shares, we will fit a linear regression, ridge regression, Bayesian linear regression model through the UCI Bike-Sharing dataset. We will compare between the linear and ridge regression models and evaluate the results obtained from Bayesian logistic regression.

Data

The data consists of 17,389 observations and the following variables:

- instant: record index
- dteday : date
- season : 1: winter, 2:spring, 3:summer, 4:fall
- yr : year (0: 2011, 1:2012)
- mnth : month (1 to 12)
- hr : hour (0 to 23)
- holiday : weather day is holiday
- weekday : day of the week
- workingday : if day is neither weekend nor holiday is 1, otherwise is 0.
+ weathersit :
- 1: Clear
- 2: Mist + Cloudy
- 3: Light Snow, Light Rain
- 4: Heavy Rain + Ice Pallets
- temp : Normalized temperature in Celsius.
- atemp: Normalized feeling temperature in Celsius.
- hum: Normalized humidity. The values are divided to 100 (max)
- windspeed: Normalized wind speed.
- casual: count of casual users
- registered: count of registered users
- cnt: count of total rental bikes including both casual and registered

Exploratory Data Analysis

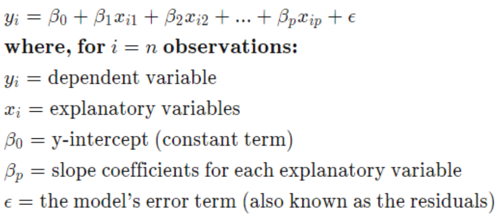
- The average normalised temperature is 0.496 and majority of normalised temperatures vary between 0.2 and 0.8
- The average normalised actual temperature is 0.47577 and majority of actual temperatures vary between 0.2 and 0.7
- The mean average normalised temperature is 0.62722 and majority of normalised temperatures vary between 0.4 and 0.9
- Most of the time, the weather situation was 1 ie. Clear, Few clouds, Partly cloudy, Partly cloudy and very less number of times the weather was Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
- The mean windspeed is 0.19 and most of time the windspeed ranges between 0.1 and 0.3
- Observations include all seasons almost equally. There is no imbalance in seasons
- Observations include all months almost equally. There is no imbalance in months
- There are almost equal number of observations for 2011 and 2012

Please refer to the Appendix for exploratory data visualisations

Methodology

Multiple Linear Regression

Multiple Linear regression attempts to model the relationship between two or more variables by fitting a **linear** equation to observed data. One variable is considered to be the dependent variable, and the others are considered to be explanatory variables. The typical structure of a multiple linear regression model is defined as follows:



$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon$

where, for $i = n$ observations:

y_i = dependent variable

x_i = explanatory variables

β_0 = y-intercept (constant term)

β_p = slope coefficients for each explanatory variable

ϵ = the model's error term (also known as the residuals)

Source: https://www.investopedia.com/terms/m/mlr

Ridge Regression

Ridge regression is a method of estimating the coefficients of multiple-regression models in scenarios where independent variables are highly correlated. Ridge regression helps to resolve the issues of over-fitting and multi-collinearity that linear regression models often suffer from. It does so by adding just enough bias to make the estimates reasonably reliable approximations to true population values, thereby reducing variance and eliminating the issues of over-fitting and multi-collinearity.

Bayesian Linear Regression

In the Bayesian viewpoint, we formulate linear regression using probability distributions rather than point estimates. The response, y, is not estimated as a single value, but is assumed to be drawn from a probability distribution. The aim of Bayesian Linear Regression is not to find the single "best" value of the model parameters, but rather to determine the posterior distribution for the model parameters. Not only is the response generated from a probability distribution, but the model parameters are assumed to come from a distribution as well. The posterior probability of the model parameters is conditional upon the training inputs and outputs:

$$P(\beta|y, X) = \frac{P(y|\beta, X) * P(\beta|X)}{P(y|X)}$$

Source: https://towardsdatascience.com/introduction-to-bayesian-linear-regression-e66e60791ea7

Here, P(β|y, X) is the posterior probability distribution of the model parameters given the inputs and outputs. This is equal to the likelihood of the data, P(y|β, X), multiplied by the prior probability of the parameters and divided by a normalization constant.

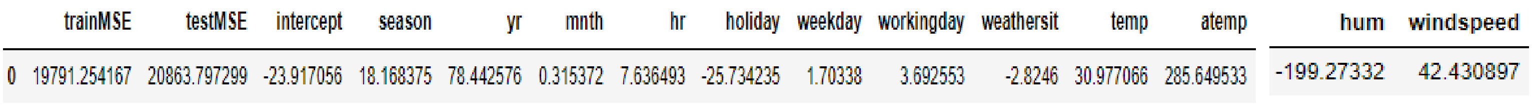
There are two primary benefits of Bayesian Linear Regression:

1. **Priors:** If we have domain knowledge, or a guess for what the model parameters should be, we can include them in our model, unlike in the frequentist approach which assumes everything there is to know about the parameters comes from the data. If we don’t have any estimates ahead of time, we can use non-informative priors for the parameters such as a normal distribution.
2. **Posterior:** The result of performing Bayesian Linear Regression is a distribution of possible model parameters based on the data and the prior. This allows us to quantify our uncertainty about the model: if we have fewer data points, the posterior distribution will be more spread out.

Results

Ridge Regression

After running the Ridge regression with lambda = e^-8, we obtain the following results:

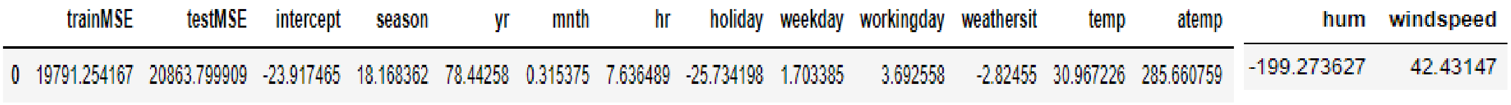


	trainMSE	testMSE	intercept	season	yr	mnth	hr	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed
0	19791.254167	20863.797299	-23.917056	18.168375	78.442576	0.315372	7.636493	-25.734235	1.70338	3.692553	-2.8246	30.977066	285.649533	-199.27332	42.430897

Image 11: Ridge regression MSE and estimated coefficients

Linear Regression

After running the linear regression, we obtain the following results:



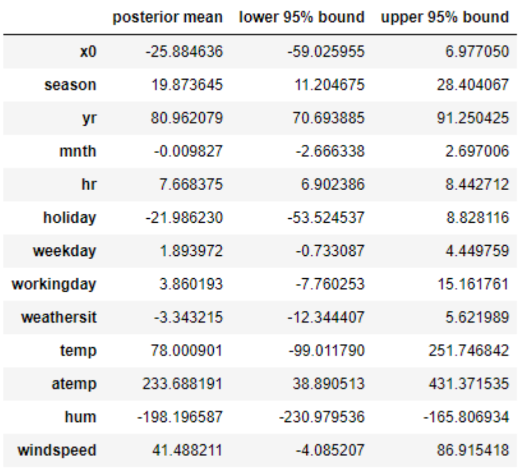
	trainMSE	testMSE	intercept	season	yr	mnth	hr	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed
0	19791.254167	20863.799909	-23.917465	18.168362	78.44258	0.315375	7.636489	-25.734198	1.703385	3.692558	-2.82455	30.967226	285.660759	-199.273627	42.43147

Image 12: Linear regression MSE and estimated coefficients

Comparing Images 11 and 12, we can see that while both models yield the same train mean squared error (MSE), the test MSE for ridge regression is slightly lesser than that for linear regression. Thus, the ridge regression model fits the data better than the linear regression model and hence it yields better results.

Bayesian Logistic Regression

After running the codes to implement Bayesian linear regression through the data, we obtain the following results:



	posterior mean	lower 95% bound	upper 95% bound
x0	-25.884636	-59.025955	6.977050
season	19.873645	11.204675	28.404067
yr	80.962079	70.693885	91.250425
mnth	-0.009827	-2.666338	2.697006
hr	7.668375	6.902386	8.442712
holiday	-21.986230	-53.524537	8.828116
weekday	1.893972	-0.733087	4.449759
workingday	3.860193	-7.760253	15.161761
weathersit	-3.343215	-12.344407	5.621989
temp	78.000901	-99.011790	251.746842
atemp	233.688191	38.890513	431.371535
hum	-198.196587	-230.979536	-165.806934
windspeed	41.488211	-4.085207	86.915418

Image 13: Results from Bayesian linear regression

From Image 14, we can infer that season, yr, hr, atemp and hum are significant variables. Month, Holiday, weekday, working day, weather situation, temperature and windspeed do not have a significant contribution in determining the sale of bikes. The season, year, hour of the day, actual feel temperature and humidity are the major determinants of sales of bikes.

Conclusion

From the above analysis, we can see that while both ridge and linear regression have the same training MSE, the test MSE for ridge regression is slightly lower. This means that the ridge regression model fits the data more efficiently than the linear regression model and thus it will yield better results. Therefore, between ridge and linear regression, we must use ridge regression to estimate sales of bike-sharing.

From the Bayesian Linear regression model, we can conclude that season, year, hour of the day, actual feel temperature and humidity are the major determinants of sales of bikes. As the season changes from 1 to 4 (ie from winter to fall), the sales of bikes also increases. This means that we can expect highest sales of bikes in summer and fall. Thus, to prepare for this high demand, companies must commission more bikes and make more pick-up/drop spots active. In contrast, companies can expect a drop in sales during winter and so lesser bikes can be commissioned to save on maintenance costs. As the year changed from 2011 to 2012, the sales of bikes also increased. This means that bike-sharing is becoming increasingly popular and companies should invest in learning more about projected sales in coming years. As actual feel temperature rises, the sales also increase. Therefore, companies can expect an increase in bike demand on warmer days. However, humidity negatively impacts sales. A unit increase in humidity can bring down bike sales by 198 units. While companies cannot control the weather temperature and humidity, it is wise for them to look at weather forecasts and commission bikes and other resources accordingly to make maximum profits.

References

Grzenda, W., 2015. The advantages of Bayesian methods over classical methods in the context of credible intervals. *Information Systems in Management*, *4*.

Pooley, C.M. and Marion, G., 2018. Bayesian model evidence as a practical alternative to deviance information criterion. *Royal Society open science*, *5*(3), p.171519.

Brownlee, J., 2020. *Imbalanced Classification with the Adult Income Dataset*. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/imbalanced-classification-with-the-adult-income-dataset/

Pham, D.T., Dimov, S.S. and Nguyen, C.D., 2005. Selection of K in K-means clustering. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, *219*(1), pp.103-119.

Bradley, P.S. and Fayyad, U.M., 1998, July. Refining initial points for k-means clustering. In *ICML* (Vol. 98, pp. 91-99).

Christopher M, n.d. Mixture Models and EM. In: *Pattern Recognition and Machine Learning*.

Tsui, A.S., Enderle, G. and Jiang, K., 2018. Income inequality in the United States: Reflections on the role of corporations. *Academy of Management Review*, *43*(1), pp.156-168.

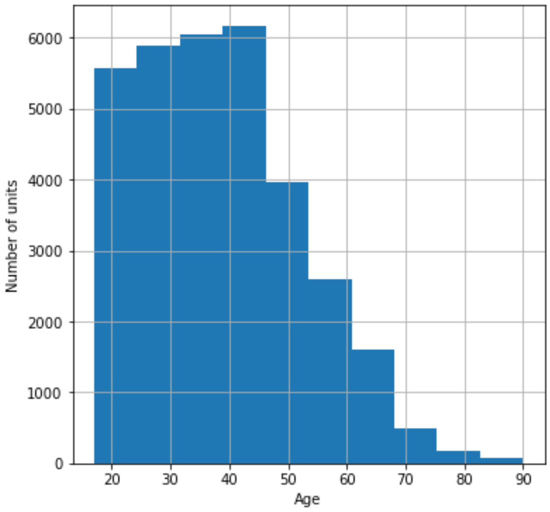
Anon,n.d. *https://cyclingindustries.com/fileadmin/content/documents/170707_Benefits_of_Bike_Sharing_UK_AI.pdf*.

Appendix

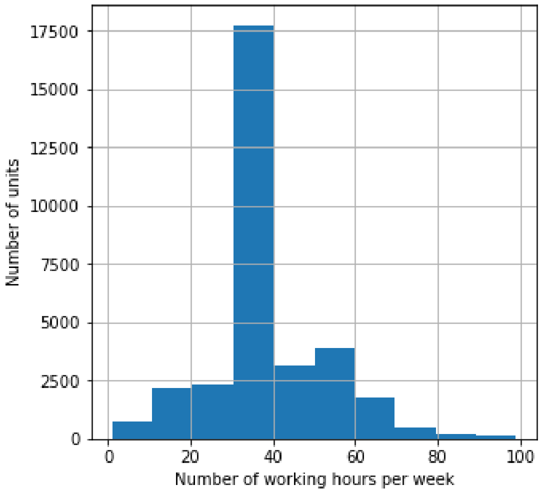
Part 1

Exploratory Data Visualisations

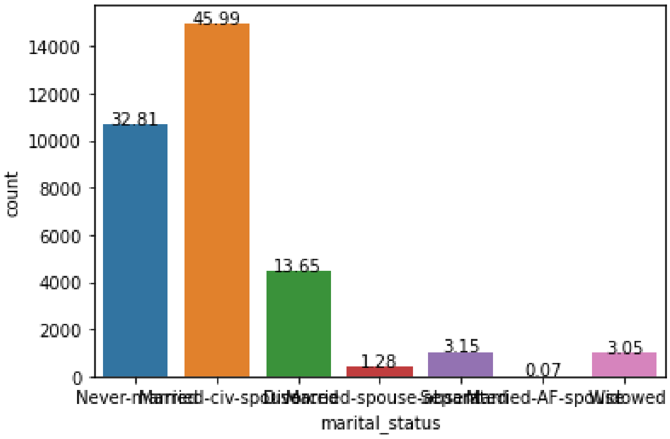
- Age distribution is right skewed and not symmetric. Most people are aged between 20 and 50. The minimum and maximum ages are 17 and 90 respectively



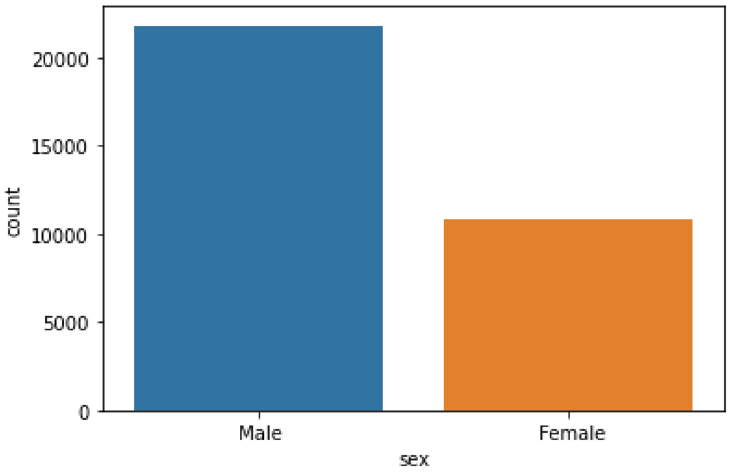
- Most people work about 30-40 hours per week



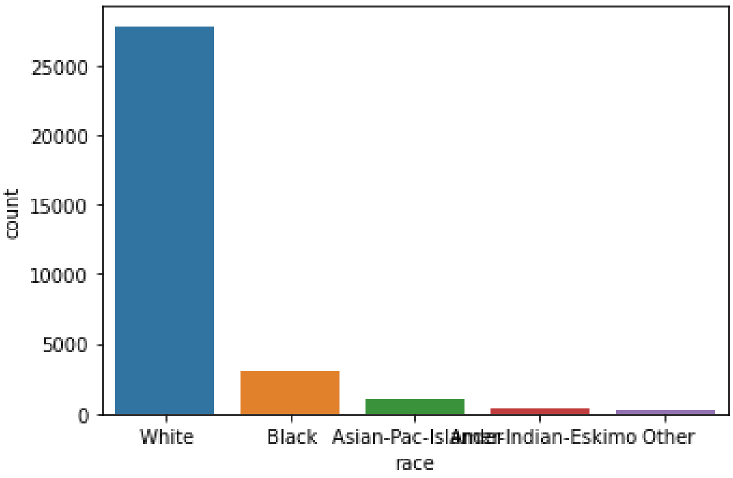
- Over 45% of the people are married



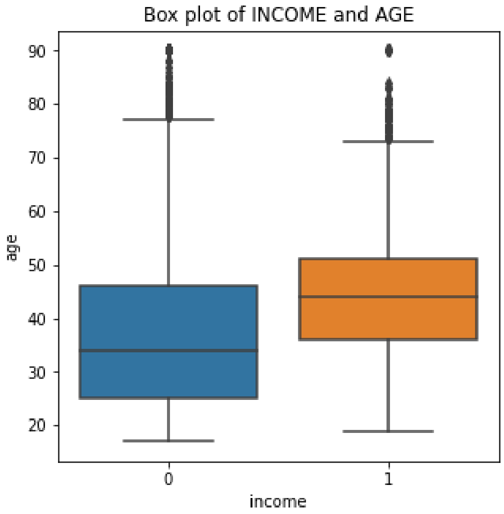
- There are more males in the data than females. The number of females are almost half the number of males



- More than half the people are white



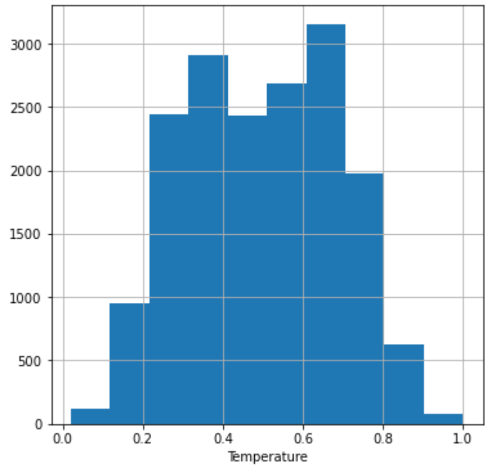
- People with income level< $50k have a median age of 34 while those with income level >=$50k is about 42



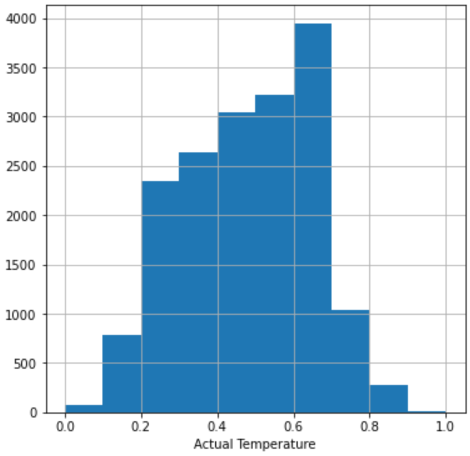
Part 2

Exploratory Data Visualisations

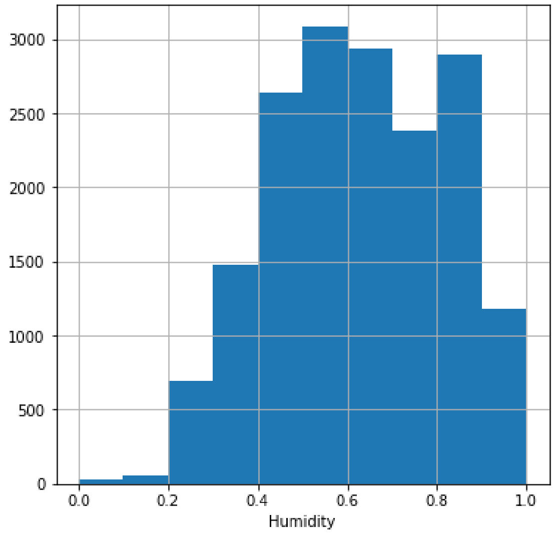
- The average normalised temperature is 0.496 and majority of normalised temperatures vary between 0.2 and 0.8



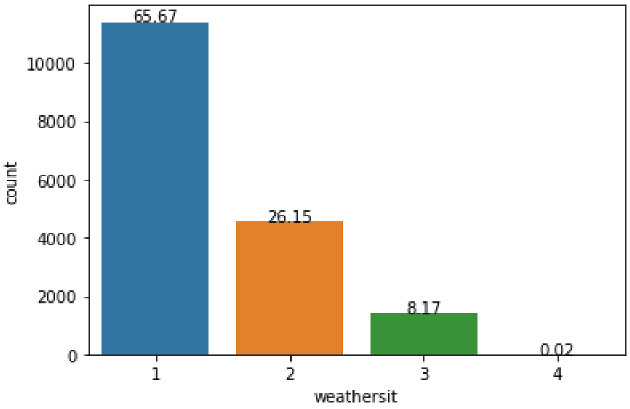
- The average actual temperature is 0.47577 and majority of actual temperatures vary between 0.2 and 0.7



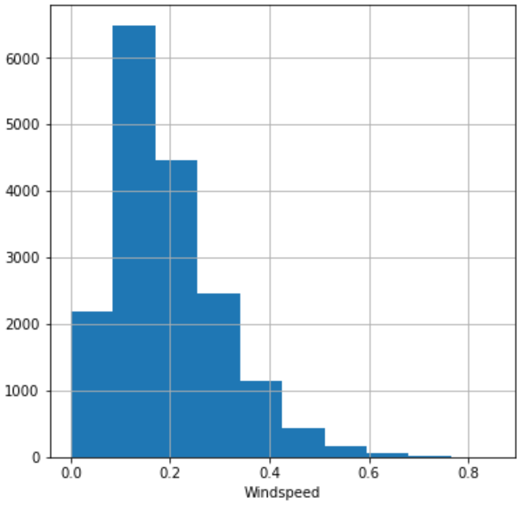
- The mean average normalised temperature is 0.62722 and majority of normalised temperatures vary between 0.4 and 0.9



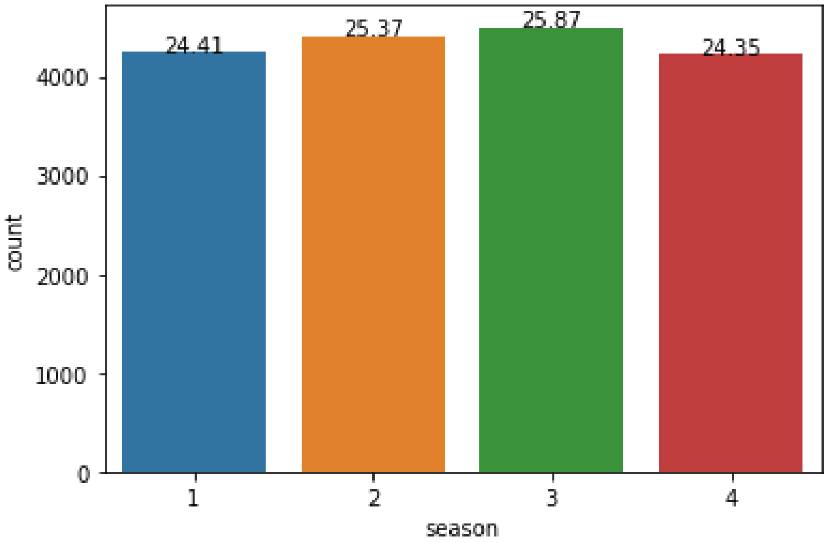
- Most of the time, the weather situation was 1 ie. Clear, Few clouds, Partly cloudy, Partly cloudy, this was followed by 2 (Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist) and 3(Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds) and very less number of times the weather was Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog



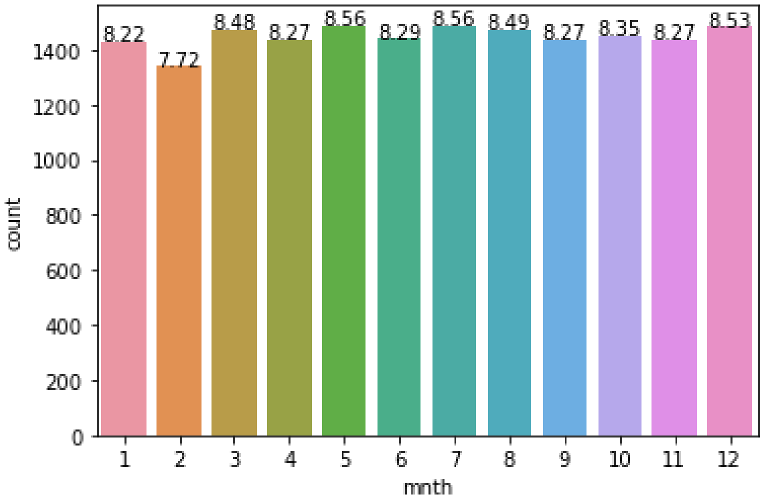
- The mean windspeed is 0.19 and most of time the windspeed ranges between 0.1 and 0.3



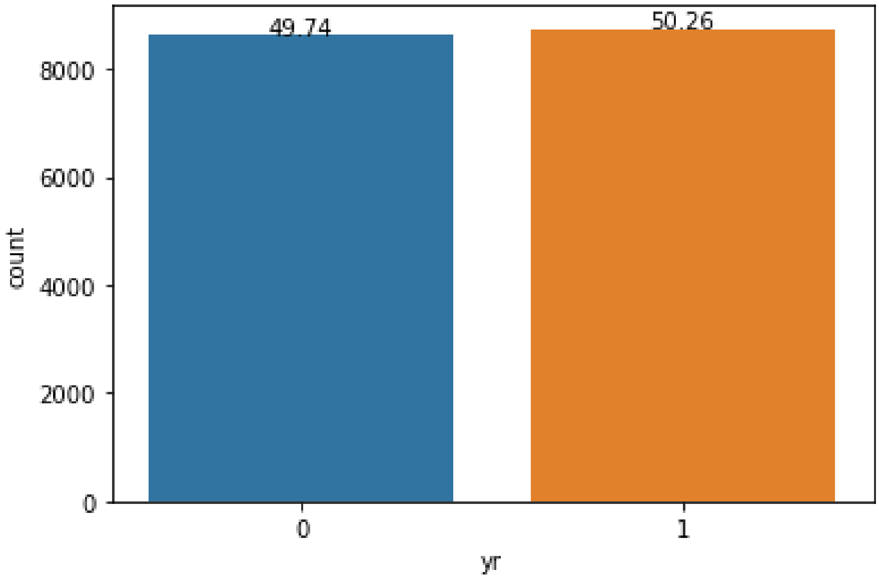
- Observations include all seasons almost equally. There is no imbalance in seasons



- Observations include all months almost equally. There is no imbalance in months



- There are almost equal number of observations for 2011 and 2012


count
yr