Danmarks Tekniske Universitet

Project 1

02450 Introduction to Machine Learning and Data Mining

Group 7

Student Contribution

Student Name - ID	Section 1	Section 2	Section 3	Exam Questions
Yu Fan Fong - s230003	40%	20%	50%	10%
Ashish Rakesh Chandra Kukreti - s230134	30%	40%	30%	20%
Karrar Adam Mahdi - s230432	30%	40%	20%	70%

1. Description of the Data Set

1.1 Overall Problem of Interest

The Bike Sharing Dataset was obtained from UCI Machine Learning Repository. The data was collated from Capital Bikeshare System based in Washington D.C. in the United States of America between 2011 and 2012 by Hadi Fanaee-T.

The dataset contains information about bike rentals, including the date and time of rental, weather conditions, and number of bikes rented. This dataset can be used for regression to predict the number of bikes rented in a given hour or day, or for classification to predict the type of day (working vs non-working).

1.2 Summarize previous analysis of the data

1. Event labeling combining ensemble detectors and background knowledge (Fanaee-T & Gama, 2014).

The paper outlines a technique for event labelling that combines ensemble detectors with prior information to increase the precision of event categorization. The bike sharing dataset is used by the authors as a case study to assess their approach. The study's objective is to show how their suggested technique may be used to categorise bike rental-related events using the attributes that were retrieved from the dataset. The suggested strategy entails creating a variety of detectors using various machine learning methods and feature sets, bagging them together, and including prior probabilities to integrate background information. In comparison to individual detectors, the authors demonstrate that the suggested strategy provides a greater level of classification accuracy for events linked to renting bicycles.

The most accurate approach identified in the study, which was used to categorise rental-related events on the dataset for bike sharing, had an accuracy rate of 85.78%. To increase the classification's accuracy, this technique combines six distinct detectors and incorporates prior knowledge regarding rental trends. The study reveals that the suggested technique for event labelling works well overall, and that it may be used for a variety of purposes other than bike sharing. The technique may be applied to enhance decision-making across a variety of fields by offering insights into the underlying patterns and trends in the data.

2. A comparative study of bike sharing demand prediction model (Zhang, 2018)

This study investigates the impact of several bike-sharing prediction models on the precision of demand forecasts. The purpose of the article is to choose the model that can most accurately forecast demand for bike-sharing. A regular time-series model, a hybrid model combining time-series and regression techniques, a deep learning model, and a hybrid model combining time-series and deep learning techniques were all compared by the author. The models were applied to the data from the Beijing bike-sharing system as part of the assessment process. According to the findings, the hybrid model combining time-series and deep learning techniques performed the best, with an average prediction accuracy of 96.48%, followed by the hybrid model combining time-series and regression methods, with an average accuracy of 95.17%.The traditional time-series model's average accuracy was 92.34%, compared to the deep learning model's average accuracy of 94.07%. This implies that combining time-series and deep learning techniques can increase the precision of projections of demand for bike-sharing.

1.3 Learning Objectives of Classification and Regression

The primary aim of applying these machine learning techniques is to identify the peak periods and weather conditions for bike demand so that bike rental operators are able to better manage their bike fleet.

- The regression model aims to **predict the total count** of bike users based on the attributes that describe the weather conditions (temp, atemp, humidity, windspeed).
- The classification model aims to **classify the type of day** (working day vs non-working day) based on the weather conditions (weathersit, temp, atemp, humidity, windspeed).

In the process, ideal weather conditions and days for high bike demand can be identified. This can be further referenced against weather forecast data to prepare the bike fleet and have a better prediction of bike demand in the near future.

1.4 Data Transformation

The dataset normalised continuous attributes like normal temp, atemp, humidity and windspeed. This step is necessary as the attributes use different scales. However, standardisation should be applied instead to use these attributes for Principal Component Analysis (PCA). The seasons and weather situation attributes should be changed to use 1-out-of-K encoding so that the difference between each season and weather situation is the same.

2. Detailed Explanation of Data Attributes

2.1 Attribute Types

Attribute Name	Attribute Type	Explanation
season	Discrete nominal	Each of the four seasons is assigned a categorical variable (1:winter, 2:spring, 3:summer, 4:fall) with no particular ranking order.
holiday	Discrete nominal	Each day is categorised as a holiday or not (0 = not holiday, 1= holiday)
weekday	Discrete nominal	Each day is given a number to indicate the day of the week (sunday = 0, saturday = 6) with no particular ranking order.
workingday	Discrete nominal	Each day is categorised as a working day or not (0 = not work day, 1 = work day), where the weekends and holidays are considered non-working days.
weathersit	Discrete nominal	The weather situation is represented using four categories as follows 1. Clear, Few clouds, Partly cloudy, Partly cloudy 2. Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist 3. Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds 4. Heavy Rain + Ice Pellets + Thunderstorm + Mist, Snow + Fog
temp	Continuous	Normalised temperature in degree Celsius

	interval	
atemp	Continuous interval	Normalised feeling temperature in degree Celsius
hum	Continuous ratio	Normalised humidity level that was originally in the range 0-100%
windspeed	Continuous ratio	Normalised wind speed values that had a max of 67mph
casual	Discrete ratio	It is represented by integers to count the number of casual (non-registered) bike users
registered	Discrete ratio	It is represented by integers to count the number of registered bike users
cnt	Discrete ratio	It is represented by integers to count the total number of bike users by summing up casual and registered users.

Note: dteday, yr, mnth and hr attributes have been omitted.

2.2 Data Issues

Missing Values

The 'missingno' Python library was used to determine the missing values in our Pandas data frame. It visualises the missing values in each column with the help of a bar graph.

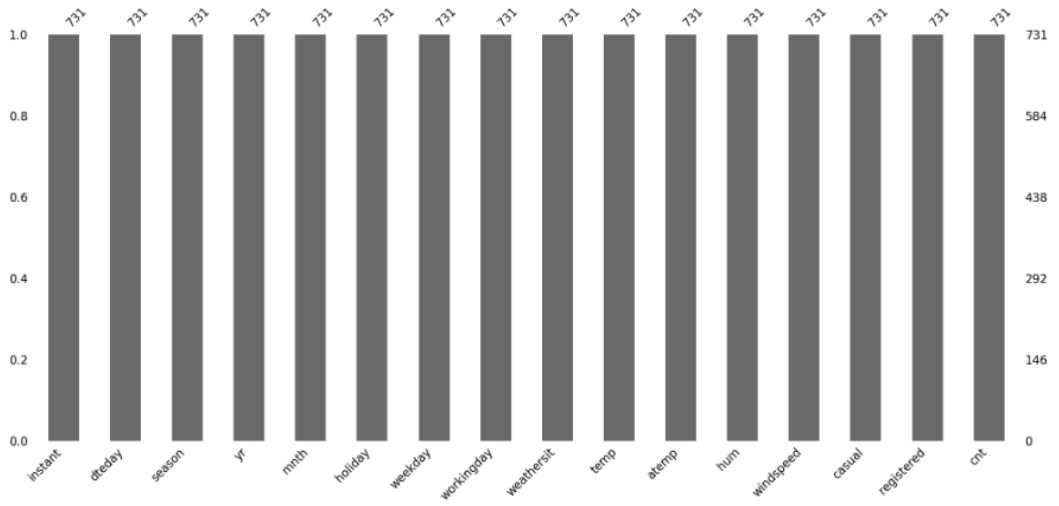

Fig. 1. Bar graph of missing values in day.csv for each attribute. The vertical axis represents the proportion of data present, with 1 meaning that 100% of the data is present with no missing values.

Outliers

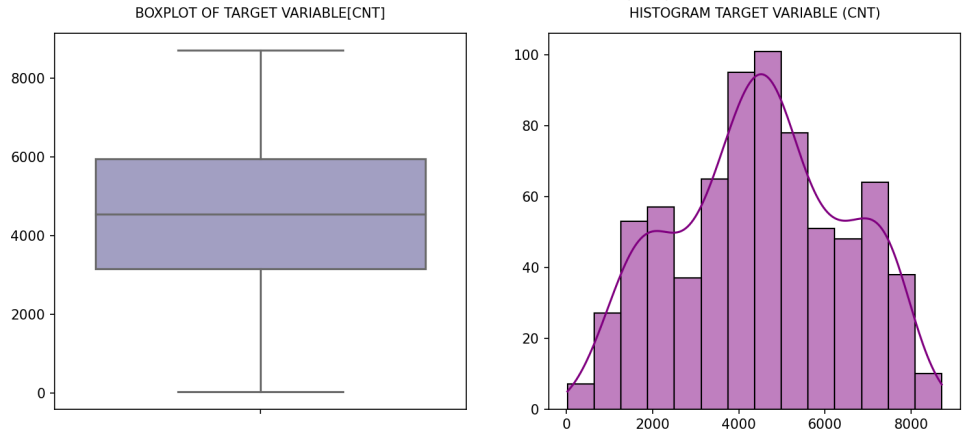

Fig. 2. Box plot of target variable (left). Histogram of target variable (right)

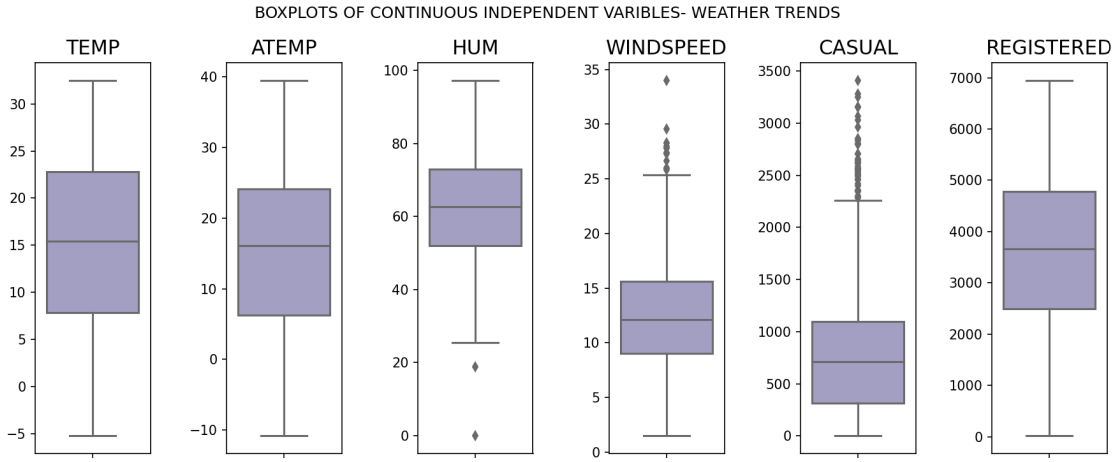

Fig. 3. Box plot of selected attributes.

From the boxplots, one of the wind speed and two of the humidity observations have values which can be considered outliers.

Although there were abnormally high wind speeds recorded, these values cannot be considered as anomalies because they correspond to the natural events occurring. For instance, there was Hurricane Sandy in October 2012, resulting in winds topping off at 67mph. One of the humidity was recorded to be 0%, which is highly unlikely. Thus, this observation will be removed.

2.3 Basic Summary Statistics

Basic statistics like median, mean, std, variance, covariance, 25th percentile, 75th percentile, min, max were calculated for day.csv's dataset for 24 months using a Python script. The box plots have been presented earlier under section 2.2.

Table 1. Basic summary statistics for selected attributes, rounded to 2 decimal places (see Appendix A for more).

Attribute	temp	atemp	hum	wind speed	casual	registered	count
Mean	15.28	15.31	62.79	12.76	848.18	3656.17	4504.35
Standard Deviation	8.60	10.76	14.24	5.12	686.62	1560.26	1937.21
Variance	74.02	115.68	202.86	26.96	471450.44	2434399.96	3752788.21

The standard deviation for atemp is slightly higher than temp, which could be due to the additional factors of wind speed and humidity that is included in the calculation of atemp.

Based on the mean values of casual and registered users, registered users take up about 80% of the total count of bike users. This suggests that there is a significant group of users who are regulars that stick to a routine and their usage is less affected by the weather conditions.

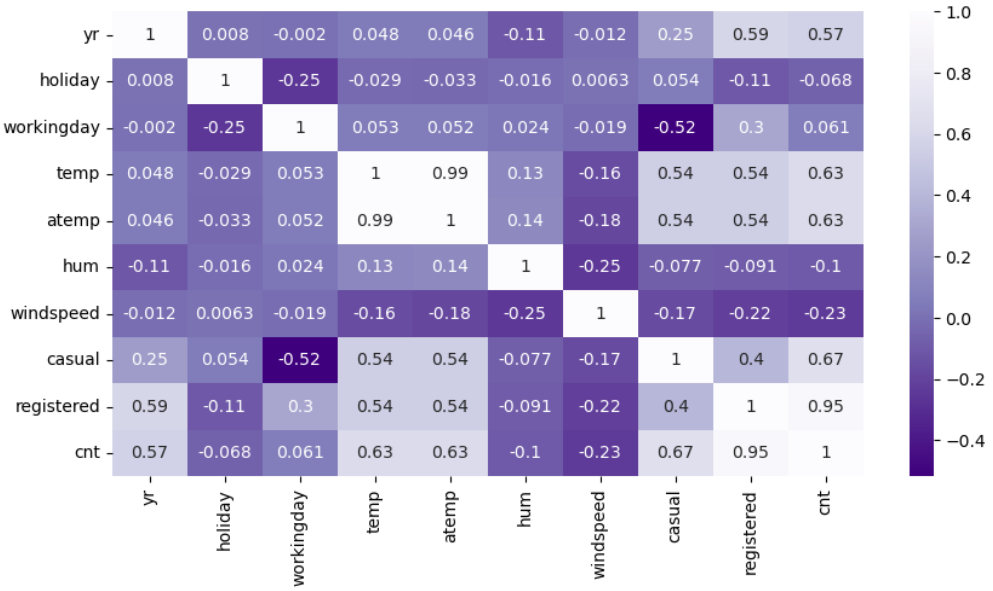

Fig 4. Covariance matrix of selected variables.

Reducing Multicollinearity

- The 'casual' and 'registered' features have positive correlations with 'cnt', indicating that they contribute to the overall bike rental count. This is expected since 'cnt' is the sum of casual and registered rentals. Since casual and registered are directly contributing to 'cnt', it is worth considering dropping them from the attributes set and only using 'cnt' as the target variable for prediction. This could simplify the model and avoid potential issues with multicollinearity.
- The 'temp' and 'atemp' features have a strong positive correlation with each other. This is expected as 'atemp' is the feeling temperature, which is typically calculated from the temp value. Since the temp and 'atemp' are measuring similar things, temp could be dropped to reduce multicollinearity in the data.

Correlation Coefficients

- The attributes 'yr', 'temp', 'atemp', 'casual' have a strong positive correlation with the 'registered' and 'cnt' attributes. This indicates that the number of registered and total bike rentals has increased over the years, suggesting a non-stationary time series.

- The 'hum' column has a negative correlation with 'registered' and 'cnt', indicating that people tend to rent fewer bikes when the humidity is high.
- The 'workingday' has a weak positive correlation with 'registered' and 'cnt', indicating that bike rentals are slightly higher on working days compared to non-working days.
- The 'windspeed' and 'holiday' attributes have weak negative correlation with 'registered' and 'cnt', indicating that people tend to rent fewer bikes on windy days or on holidays.
- Overall, it seems that temperature and humidity have the strongest correlations with bike rentals, while the other features have weaker correlations.

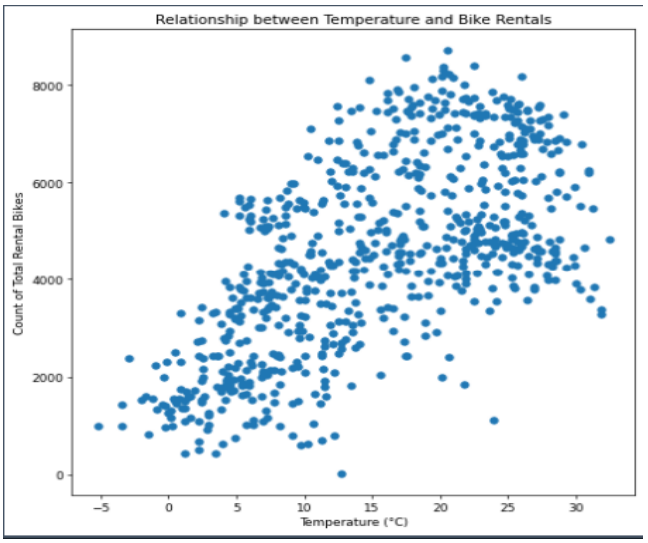

Fig. 5. Scatter plot of total count against temperature.

As an example, the scatter plot in figure 5 shows a positive correlation between temperature and bike rentals, where warmer temperatures tend to result in higher bike rentals. In other words, as the temperature increases, the number of bike rentals is also increasing.

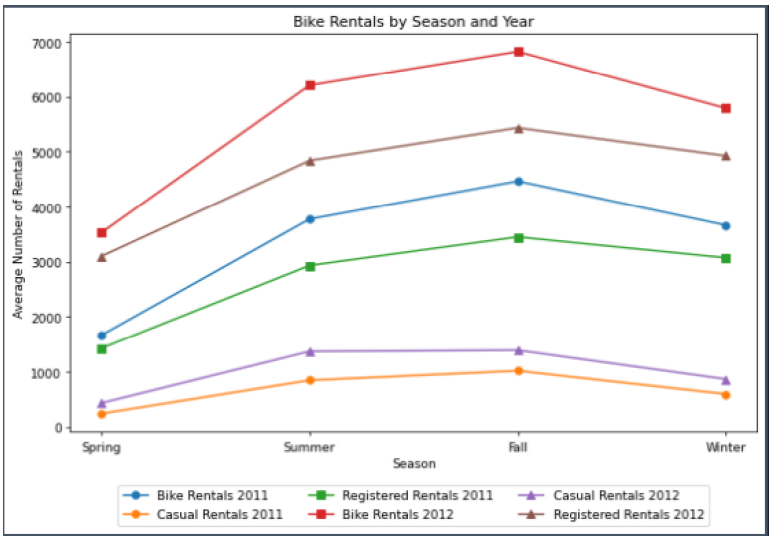

Fig. 6. Line plot of bike users against season.

From figure 6, there appears to be a seasonal demand over the two years, with bike usage lowest in the winter, and increasing until fall the following year.

3. Data Visualisation and PCA

3.1 Variance Explained

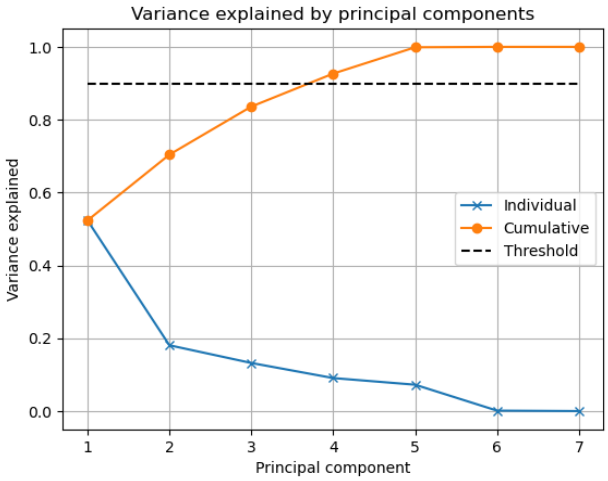

Fig. 7. Line of variance explained against PCs.

From figure 7, it is evident that more than 90% of the variance can be explained by the first four principal components.

3.2 Principal Directions of the Considered PCA Components

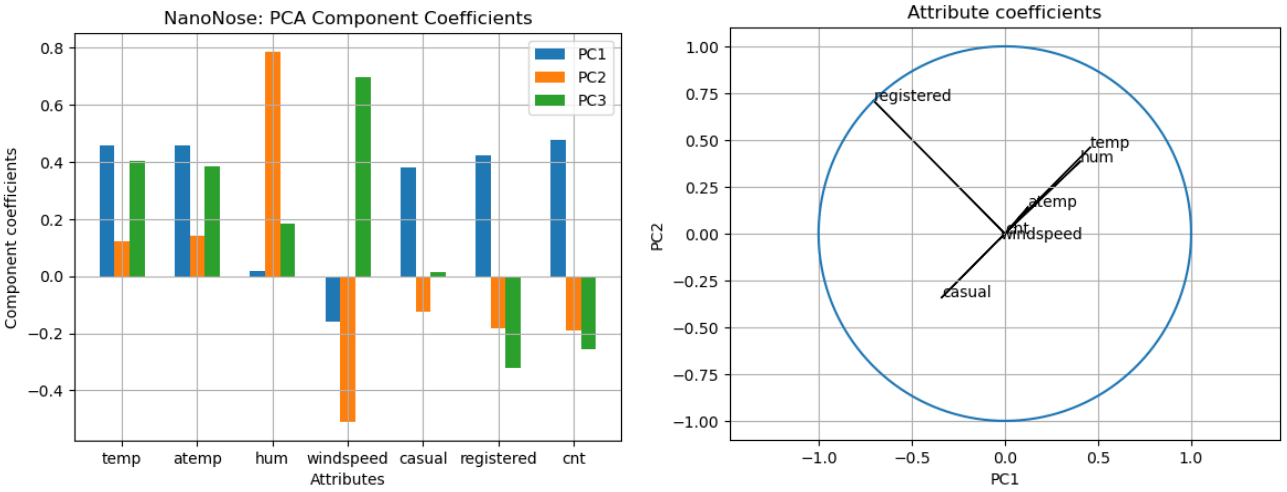

Fig. 8. Bar graph of PC loadings for the first three PCs (left). Vector plot of PC1 loadings against PC2 loadings (right).

Interpreting principal directions of PC1, PC2 and PC3 in terms of the features

PC1

- Days with with high temperatures and users, low wind speed → High PC1 values
- Days with with lower temperatures and users, high wind speed → Low PC1 values

PC2

- Days with with high temperatures and wind speed, low users → High PC1 values
- Days with with lower temperatures and wind speed, high users → Low PC1 values

PC3

- Days with with high humidity, low wind speed → High PC1 values
- Days with with lower humidity, high wind speed → Low PC1 values

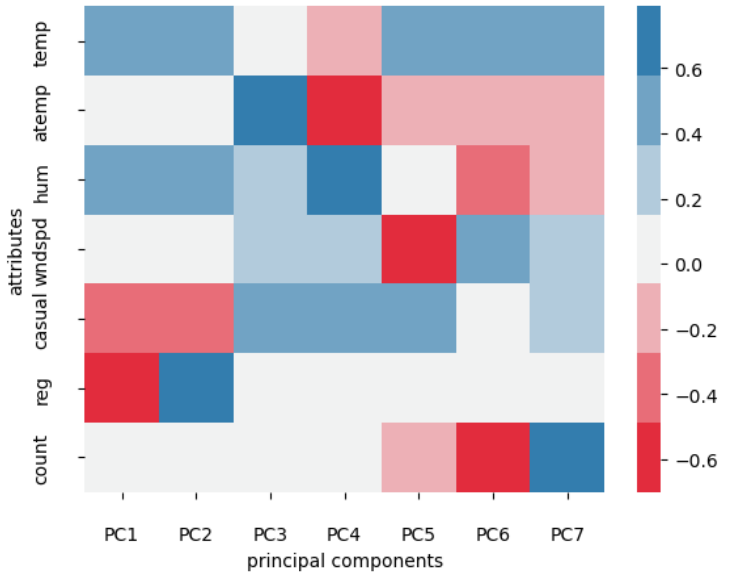

Fig. 9. Heat map of PC loadings (see Appendix B for numerical values).

3.3 Data Projected onto the Considered Principal Components.

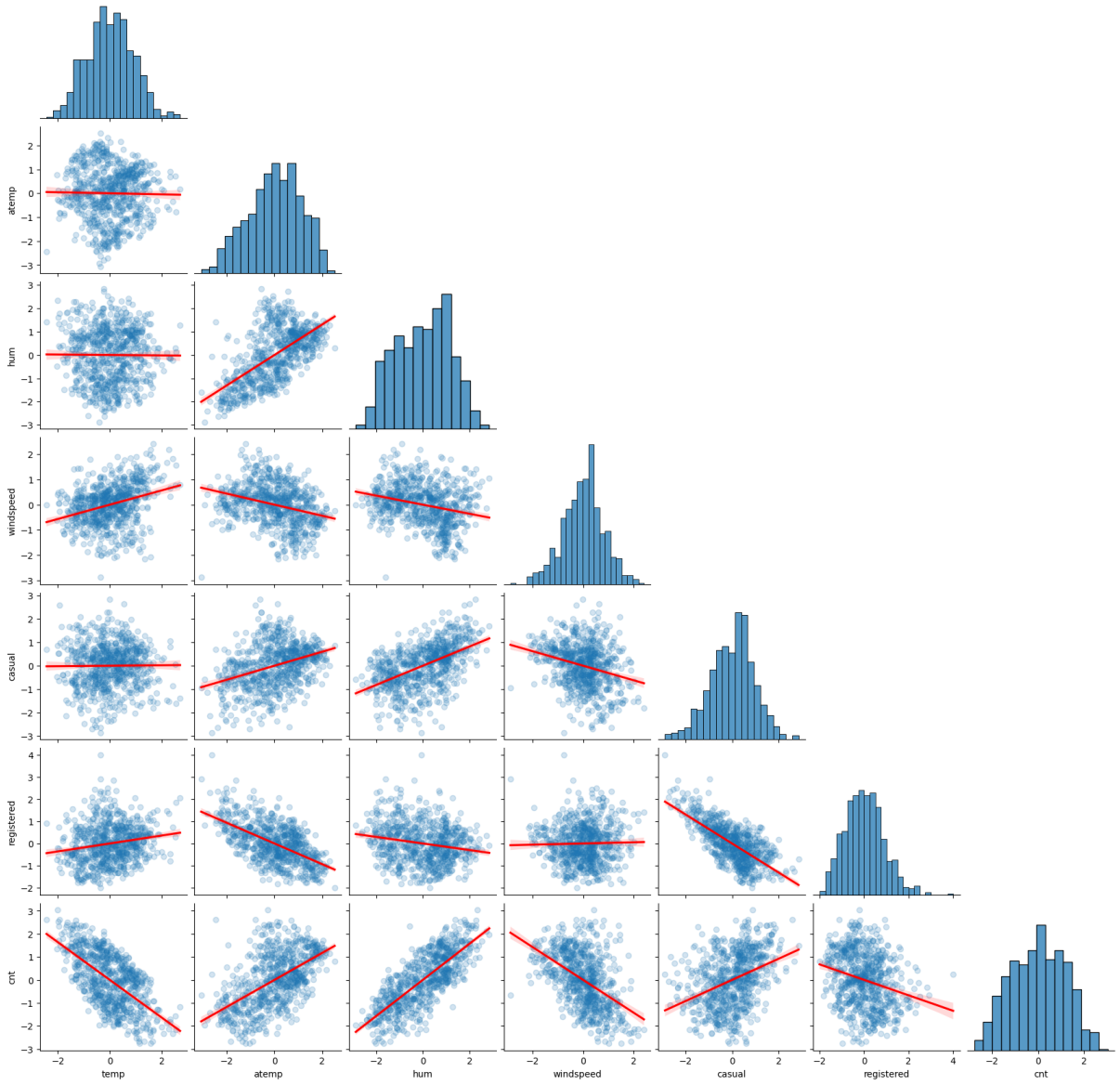

Fig. 10. Pair plot of projected data.

4. Discussion on Learning Points from Data

The dataset provided is rather complete and of high quality, with minor issues. PCA showed that it is possible to reduce the dimensionality of the dataset as close to 100% of the variance can be explained using five PCs.

Thus, a simple linear regression model should be sufficient to predict the total count of bike users based on the weather conditions.

5. Exam Problems

Question 1. Spring 2019 question 1:

Correct Answer is : D x1 (Time of day) is interval, x6 (Traffic lights) is ratio, x7 (Running over) is ratio, and y (Congestion level) is ordinal.

X1: As mentioned in the question the right of the attribute is coded into a 30 minutes interval, which is an interval category.

Both X6 (traffic light) and X7(Running over) are rationed since they represent physical objects, and have significant differences between values, so both X6 (traffic light) and X7(Running over) are rationed data with clear order values.

Y(congestional level) is ordinal : congestion level as it present in the dataset has rank between values as proposed "low to high" lead us to the conclusion it is Ordinal data.

Question 2. Spring 2019 question 2

Correct Answer is : A = 7.0

Max-norm distance $(p, \infty) \Rightarrow d\infty(x,y) = \max\{/x_1 - y_1/ , /x_2 - y_2/, ..., /x_m - y_m/$

For $p = \infty$, $dp(x_{14}, x_{18}) = \max\{/26 - 9/ , /0 - 0/, /2 - 0/ ...., /0 - 0/ = 7.0$

Question 3.Spring 2019 question 3

Correct Answer is : C: The variance explained by the first two principal components is less than 0.5

The singular values are given in the diagonal matrix ***S.*** the total variance is the sum of diagonal square in matrix ***S***

The variance can be illustrated by each principal component being given,by the proportion of its eigenvalue to the total variance. variance explained by PC1 $= (0.49)^2 + (0.58)^2 + (0.56)^2 + (0.31)^2 + (-0.06)^2 = 0.6913$

PC2 = 0.2433, PC3 = 0.1441, PC4 = 0.1103, PC5 = 0.1030

In the given SVD decomposition, the diagonal matrix S contains eigenvalues of the covariance matrix, therefore the total variance of the data can be calculated as the sum of the eigenvalues . In this case the total variance is the sum of all eigenvalues of S as :

The total variance : 13.9 + 12.47 + 11.48 + 10.03 + 9.45 = 57.33

The proportion of variance can be seen by the first two principal components , we sum the first two eigenvalues in S and divide them by the total variance:

(13.9 + 12.47)/57.33 ≈ 0.450 which is less than 0.5, so C is the correct answer.

Question 4.Spring 2019 question 4

Correct Answer is : D as stated an observation with a low value of time of day, high value of broken truck, a high of Accident victim, and high value of detects will typically positive value of projection onto principal component number 2

Question 5: Spring 2019 question 14

Correct Answer is : B : Jaccard similarity of s_1 and s_2 is 0.000650

The Jaccard similarity would be low, as the two documents only share the word "the". The Jaccard similarity would be:

Jaccard similarity = |s1 ∩ s2| / |s1 ∪ s2| = |{"the"}| / |{"the", "bag", "of", "words", "representation", "becomes", "less", "parsimonious", "if", "we", "do", "not", "stem", "words"}| = 1 / 13 = 0.000650.

Question 6: Spring 2019 question 27

Correct Answer is : $\mathbf{P}(\widehat{x_2} = 0 \mid y = 2) = 0.116$

References

Fanaee-T, Hadi. (2013). Bike Sharing Dataset. UCI Machine Learning Repository. https://doi.org/10.24432/C5W894.

Fanaee-T, H., & Gama, J. (2014). Event labeling combining ensemble detectors and background knowledge. In M. Soares, F. Pereira, H. T. Nguyen, M. Detyniecki, & R. Meo (Eds.), Proceedings of the 2014th European Conference on Machine Learning and Knowledge Discovery in Databases (ECML/PKDD 2014) (pp. 433-448). Springer.

Zhang, Y. (2018). A comparative study of bike sharing demand prediction model. In X. Huang, Y. Fang, S. Zhu, & S. Liu (Eds.), Proceedings of the 2018 International Conference on Artificial Intelligence and Big Data (ICAIBD 2018) (pp. 115-119). Association for Computing Machinery (ACM).

Appendix A

Full table of summary statistics

Attribute	Mean	Standard Deviation	Variance	Median	25th percentile	75th percentile	Max	Min
temp	15.28	8.60	74.02	15.42	7.84	22.80	32.50	-5.22
atemp	15.31	10.76	115.68	16.12	6.30	24.17	39.50	-10.78
hum	62.79	14.24	202.86	62.67	52.00	73.02	97.25	0.0
Wind speed	12.76	5.12	26.96	12.13	9.04	15.63	34.00	1.50
Casual	848.18	686.62	471450.44	713.00	315.50	1096.00	3410.00	2.00
Registered	3656.17	1560.26	2434399.96	3662.00	2497.00	4776.50	6946.00	20.00
Count	4504.35	1937.21	3752788.21	4548.00	3152.00	5956.00	8714.00	22.00

Appendix B

Numerical values of PCA Loadings

$$\begin{bmatrix} 0.46 & 0.46 & 0.02 & -0.16 & 0.38 & 0.42 & 0.48 \\ 0.12 & 0.14 & 0.79 & -0.51 & -0.13 & -0.18 & -0.19 \\ 0.4 & 0.38 & 0.18 & 0.7 & 0.02 & -0.32 & -0.25 \\ -0.03 & -0.03 & 0.28 & 0.31 & -0.69 & 0.56 & 0.2 \\ -0.34 & -0.34 & 0.52 & 0.37 & 0.54 & 0.08 & 0.26 \\ -0.7 & 0.71 & -0.01 & 0.02 & 0.0 & -0.0 & -0.0 \\ 0.0 & -0.0 & 0.0 & 0.0 & -0.27 & -0.6 & 0.75 \end{bmatrix}$$