UNIVERSITY OF MORATUWA

FACULTY OF ENGINEERING

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

MBA/PG Diploma in Information Technology
2015 Intake Semester 3 Examination

CS5122 DESCRIPTIVE AND PREDICTIVE ANALYTICS

Time allowed: 2 Hours December 2015

ADDITIONAL MATERIAL: *None*

INSTRUCTIONS TO CANDIDATES:

1. This paper consists of **5** questions in **8** pages.
2. Answer **all** questions.
3. Color figures are given separately as Annex.
4. Start answering each of the main questions on a new page.
5. The maximum attainable mark for each question is given in brackets.
6. This examination accounts for 40% of the module assessment.
7. This is a closed book examination.

   NB: It is an offence to be in possession of unauthorised material during the examination.
8. Only calculators approved by the Faculty of Engineering are permitted.
9. Assume reasonable values for any data not given in or with the examination paper. Clearly state such assumptions made on the script.
10. In case of any doubt as to the interpretation of the wording of a question, make suitable assumptions and clearly state them on the script.
11. This paper should be answered only in English.

Question 1 (20 marks)

Answer the following questions based on the given Descriptive Statistics related to a person’s IQ score (IQ) and his/her brain size, height, and weight.

	IQ Score	Brain Size	Height	Weight
	124	81.69	64.5	118
	150	103.84	73.3	143
	128	96.54	68.8	172
	134	95.15	65	147
	124	81.69	64.5	118
	…	…	…	…
	81	83.43	66.5	143
	128	94.81	66.5	153
	124	94.94	70.5	144
	94	89.4	64.5	139
	74	93	74	148
N	38	38	38	38
Mean	111.34	90.67	68.42	151.05
Median	115	90.54	68	146.5
Mode	124	#N/A	64.5	118
Variance	510.66	52.65	15.95	551.24
Skewness	-0.116	0.414	0.545	0.135
Kurtosis	-1.278	-0.156	-0.384	-0.934
Min	72	79.06	62	106
Max	150	107.95	77	192

Data source: Willerman, et al, 1991

(i) Is there a considerable variation in IQ Score across different people? Comment while considering relevant statistics. [2]

(ii) What does the difference between the mean, median, and mode tell you? [3]

(iii) Comment on the shape of the distributions for both IQ Score and Brain Size. [4]

(iv) Following Correlation coefficients are observed: [4]

	Brain Size	Height	Weight
IQ Score	0.3778	-0.0932	0.0025
Brain Size		0.5884	0.5135
Height			0.6996

How would you interpret these observations? Explain.

(v) Propose a suitable technique to predict the IQ Score of a person given his/her Brain Size, Height, and Weight. Justify. [3]

(v) Empirical CDFs of IQ Score and Brain Size are given in Figure 1 (see Annex). What can you claim from those graphs? Discuss while justifying your claims. [4]

Question 2 (20 marks)

(i) Using a suitable real-world example, describe how Collaborative Filtering is used in prediction. [5]

(ii) The following results were obtained by performing PCA on a 12 variable dataset related to people. Scatter plot in Figure 2 shows PC1 and PC2. Four clusters indicate different classes of people.

```
> summary(people.pca)
  Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6
Standard deviation     2.5357 1.4975 1.2719 0.99899 0.56450 0.40648
Proportion of Variance 0.5358 0.1869 0.1348 0.08317 0.02655 0.01377
Cumulative Proportion  0.5358 0.7227 0.8575 0.94066 0.96722 0.98098

                           PC7     PC8     PC9    PC10    PC11    PC12
                       0.31527 0.23451 0.16568 0.15537 0.12329 0.08371
                       0.00828 0.00458 0.00229 0.00201 0.00127 0.00058
                       0.98927 0.99385 0.99614 0.99815 0.99942 1.00000
```

(a) How many Principle Components are suitable to represent this dataset? Justify your answer. [4]

(b) Based on the PCA outputs what can you conclude about the data? Discuss your answer. [4]

(c) Briefly describe a how a new person (that is not included in the dataset) can be classified based on these findings. [3]

(iii) Discuss how outliers and variables of different scales may impact the PCA. [4]

Question 3 (20 marks)

(i) Which clustering technique would you recommend to cluster each of the datasets shown in Figure 3? Justify your answer. [3 × 3]

(ii) Using one of the graphs in Figure 3 as an example, explain how the *k*-Nearest Neighbor Classification technique can be used to classify a new data point. [5]

(iii) Figure 4 shows a cluster dendogram of cities. What can you conclude from this cluster analysis? Discuss. [6]

Question 4 (20 marks)

(i) Scatter plot matrix in Figure 5 shows the relationship between a demand for heating oil vs. price of heating oil and family income. Multiple Linear Regression model parameters are also shown below.

```
lm(d$Demanded ~ d$Price + d$Income)

Coefficients:
(Intercept)      d$Price     d$Income
    -2.1050      -0.5788       4.0750
```

(a) Write the corresponding Multiple Linear Regression equation. [2]

(b) By analyzing the scatter plots and regression model parameters what can you claim about the relationship between demand, price, and income? Justify your claims. [6]

(c) Predict the heating oil demand from a family, where price = 80 and Income = 15. [3]

(d) Discuss the accuracy of the fitted Multiple Linear Regression model. [4]

(ii) Figure 6 visualizes the Salary Distribution by Age Bins. Discuss what you can conclude from the graph. [5]

Question 5 (20 marks)

(i) Line charts in Figure 7 shows the number of fans added Laura Marling (a singer) via Facebook and YouTube with time.

(a) Discuss what can be learned from these time series while considering the Trend, Seasonal, Cyclical, and Irregular time series components. [6]

(b) What time series prediction model would you recommend to predict the total number of fans for the next month starting December 2011? Justify your recommendation. [5]

(c) Is it better to predict the “Fans - Total” time series directly or predict it by predicting “Fans - Facebook” and “Fans - YouTube” time series separately? Discuss. [3]

(ii) 2 graphs in Figure 8 present time series data about hurricanes in an alternative form. Discuss how the graphs should be interpreted and what can we conclude from the graph. [6]

---- END OF PAPER ---

Annex – Figures

Question 1

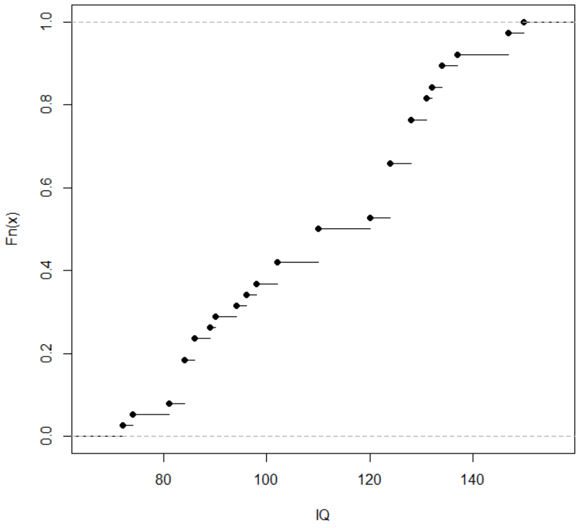


Figure 1(a) – IQ

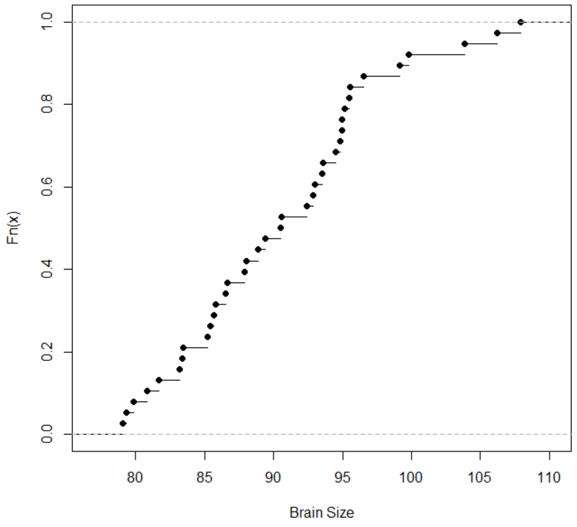


Figure 1(b) – Brain Size

Question 2

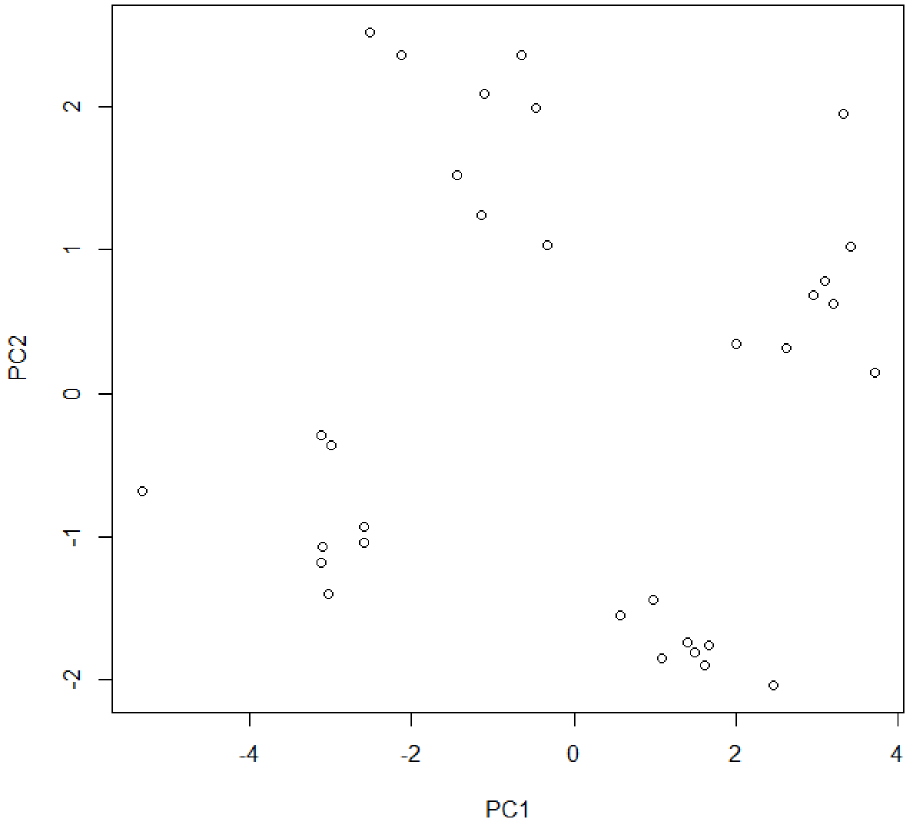


Figure 2

Question 3

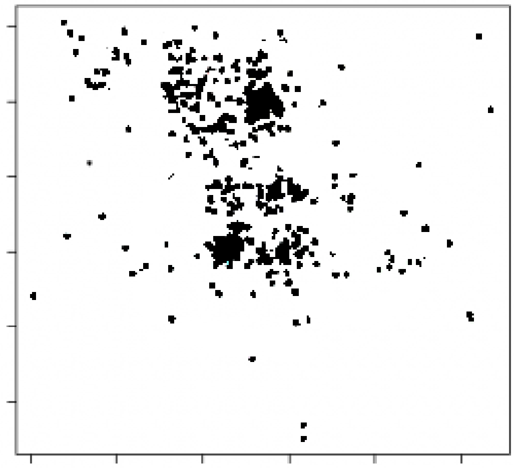


Figure 3(a)

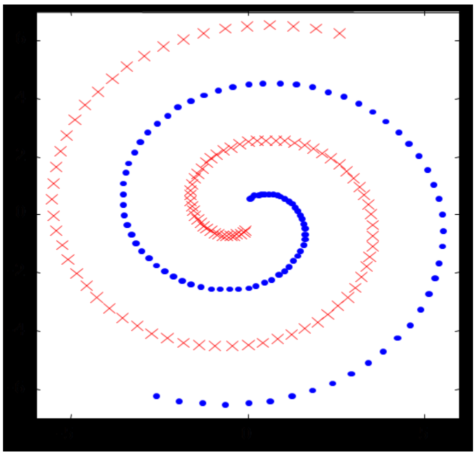


Figure 3(b)

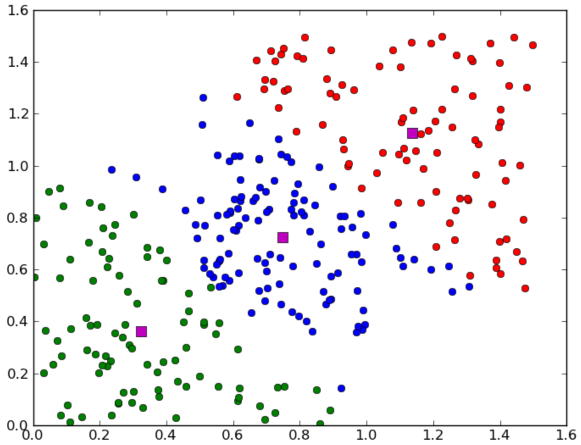


Figure 3(c)

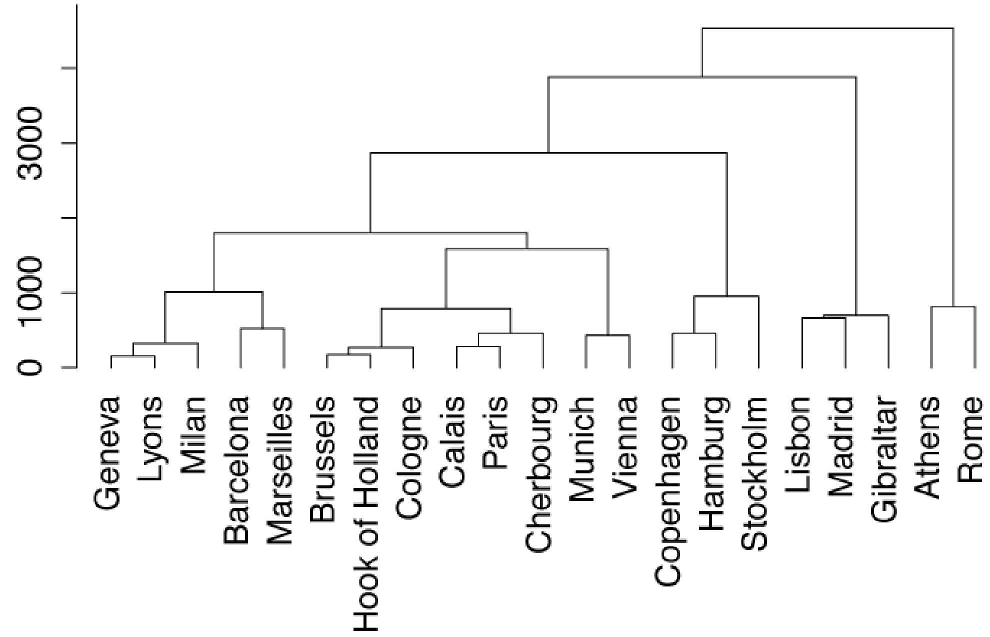


Figure 4

Question 4

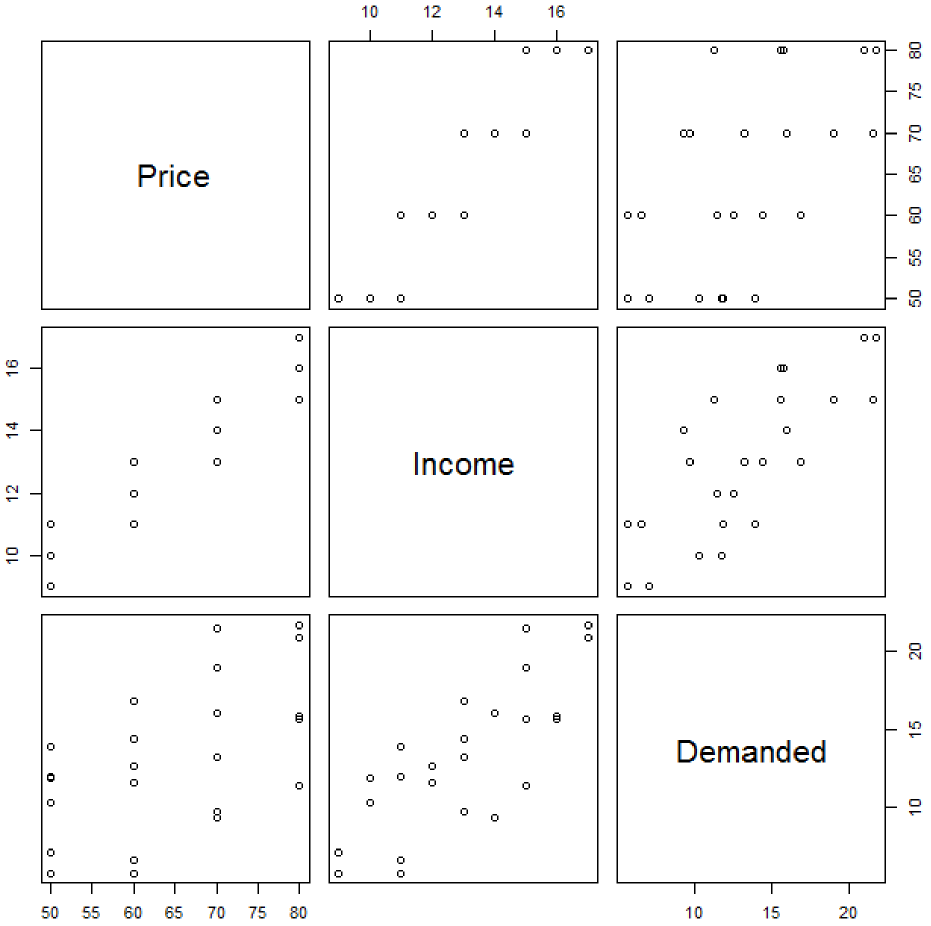


Figure 5

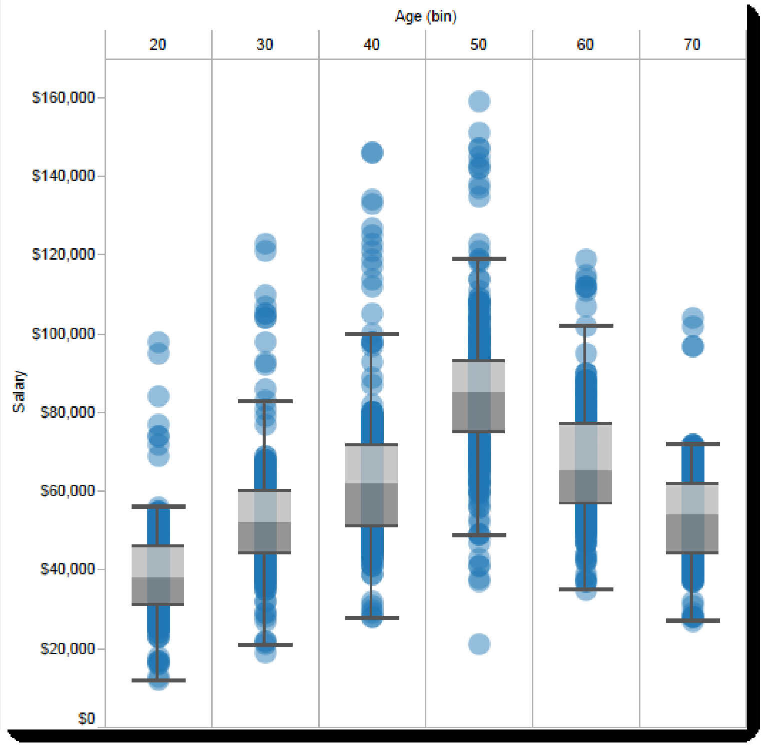
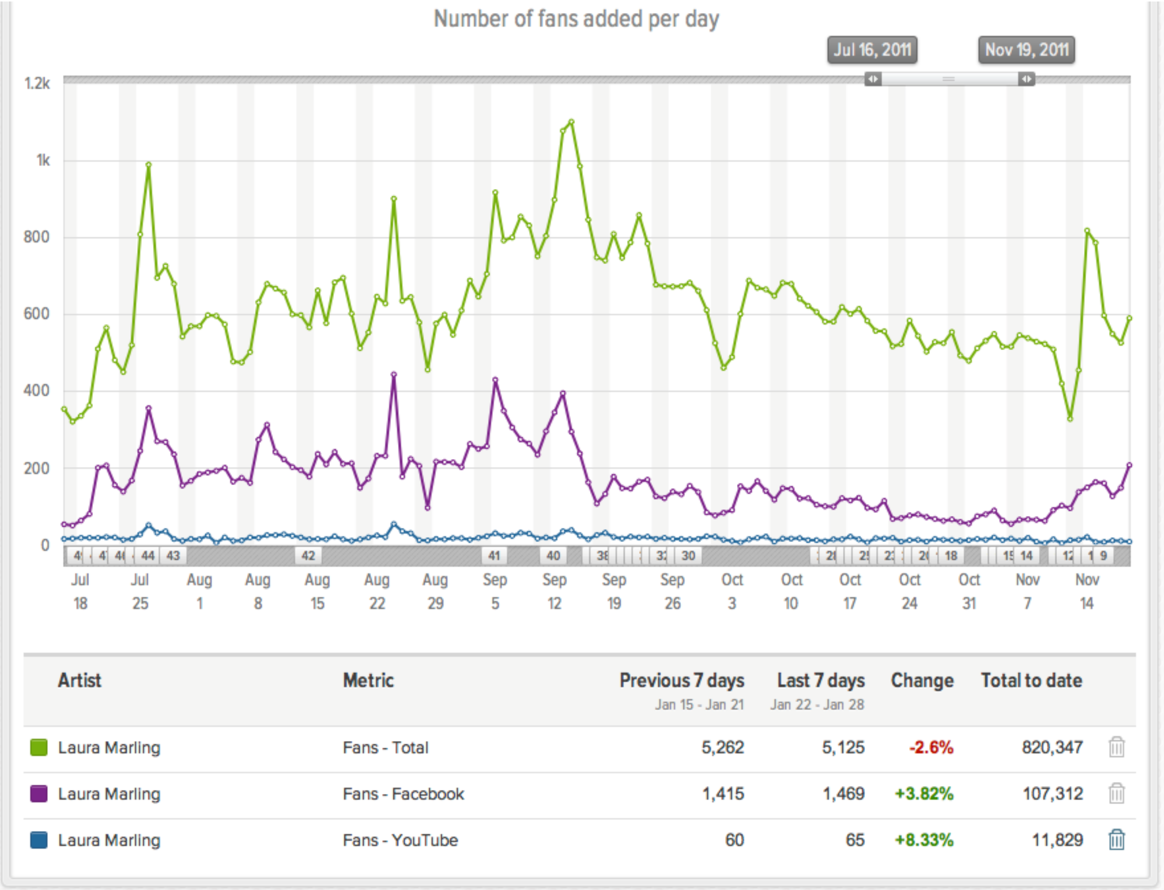


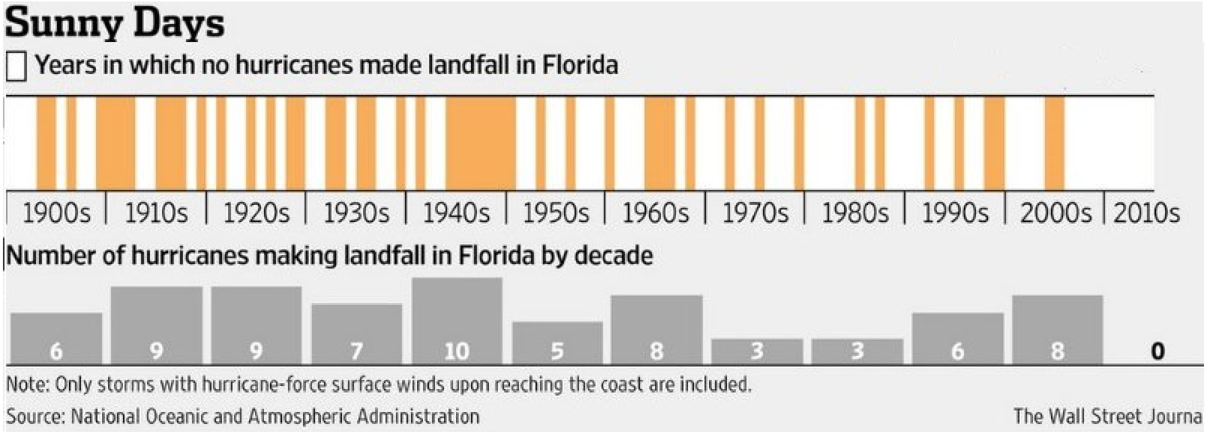
Figure 6



Artist	Metric	Previous 7 days Jan 15 - Jan 21	Last 7 days Jan 22 - Jan 28	Change	Total to date
🟩 Laura Marling	Fans - Total	5,262	5,125	-2.6%	820,347
🟪 Laura Marling	Fans - Facebook	1,415	1,469	+3.82%	107,312
🟦 Laura Marling	Fans - YouTube	60	65	+8.33%	11,829

Source: http://knowledgebase.musicmetric.com

Figure 7


Sunny Days

□ Years in which no hurricanes made landfall in Florida

Number of hurricanes making landfall in Florida by decade

Note: Only storms with hurricane-force surface winds upon reaching the coast are included.

Source: National Oceanic and Atmospheric Administration

The Wall Street Journa

Figure 8