# UNIVERSITY OF MORATUWA

## FACULTY OF ENGINEERING

## DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

MBA/PG Diploma in Information Technology
2016 Intake Semester 4 Examination

## CS5122 DESCRIPTIVE AND PREDICTIVE ANALYTICS

Time allowed: 2 Hours April 2017

---

**ADDITIONAL MATERIAL:** *None*

**INSTRUCTIONS TO CANDIDATES:**

1. This paper consists of **5** questions in **8** pages.
2. Answer **all** questions.
3. 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 salaries of a set of employees and cost of MBA programs in the USA as of 2011.

| | Salary | Age | MBA | | School | Total MBA Cost | 2-Year Tuition |
|---|---|---|---|---|---|---|---|
| | $ 28,260 | 25 | No | | Columbia | $168,307 | $106,416 |
| | $ 43,392 | 28 | Yes | | Pennsylvania | $168,000 | $108,018 |
| | $ 56,322 | 37 | Yes | | Stanford | $166,812 | $106,236 |
| | $ 26,086 | 23 | No | | Chicago | $165,190 | $101,800 |
| | $ 36,807 | 32 | No | | Dartmouth | $162,750 | $101,400 |
| | $ 57,119 | 57 | No | | MIT | $160,378 | $100,706 |
| | $ 48,907 | 45 | No | | Harvard | $158,800 | $97,200 |
| | $ 34,301 | 32 | No | | New York | $157,622 | $94,572 |
| | $ 31,104 | 25 | No | | Northwestern | $156,990 | $102,990 |
| | $ 60,054 | 57 | No | | Yale | $151,982 | $99,800 |
| | $ 41,420 | 42 | No | | Carnegie Mellon | $149,400 | $105,000 |
| | | | | | | | |
| **N** | 35 | 35 | Yes = 6<br>No = 29 | | | 20 | 20 |
| **Mean** | $45,310 | 40 | N/A | | | $ 150,795 | $99,328 |
| **Median** | $42,377 | 37 | N/A | | | $ 150,691 | $99,400 |
| **Mode** | N/A | 32 | N/A | | | N/A | N/A |
| **Variance** | 172,606,079 | 150.46 | N/A | | | 162,471,581 | 36,279,634 |
| **Skewness** | 0.7822 | 0.3246 | N/A | | | -0.228 | -0.56382 |
| **Kurtosis** | 0.7054 | -1.3348 | N/A | | | -1.0635 | 0.3113 |
| **Min** | $ 26,086 | 23 | N/A | | | $ 127,144 | $84,800 |
| **Max** | $ 84,876 | 61 | N/A | | | $ 168,307 | $108,018 |

Source: Introduction to Applied Multivariate Analysis with R by B. Everitt and T. Hothorn, Springer and Schools.

(i) Calculate the following statistics:

(a) Standard Deviation of Salary [2]

(b) Range of 2-Year Tuition Cost [2]

(ii) Comment on the shape of the distributions for both Salary and Age. [4]

(iii) The correlation between Salary and Age is 0.8814. What can we conclude from this? [2]

(iv) Following table presents statistics for employees with and without an MBA: [4]

| Salary | MBA = Yes | MBA = No |
|---|---|---|
| **Mean** | $51,510 | $44,028 |
| **Variance** | 309,360,803 | 144,410,167 |

What can you conclude based on these statistics? Discuss.

(v) Given these statistics does it make sense to have an MBA? Discuss. [6]

## Question 2 (20 marks)

(i) What is mean by "the curse of dimensionality"? [2]

(iii) Using an example, explain how PCA (Principle Component Analysis) can help a data analyst to overcome the curse of dimensionality. [3]

(iii) The following results were obtained by performing PCA on an 11 variable dataset related to cities.

The dataset contains attributes such as area of city, population (1980, 1990, and 2000), growth, resource consumption (e.g., Food, Water, Electricity, and Phones), and number of vehicles. See Scatter plot matrix in Fig. 1.

```
## Importance of components:
##                           Comp.1    Comp.2     Comp.3     Comp.4     Comp.5     Comp.6     Comp.7
## Standard deviation     2.4629420 1.3151570 1.00246883 0.92086602 0.83523273 0.50012846 0.48128911
## Proportion of Variance 0.5514621 0.1572398 0.09135852 0.07709038 0.06341943 0.02273895 0.02105811
## Cumulative Proportion  0.5514621 0.7087019 0.80006045 0.87715083 0.94057026 0.96330921 0.98436732
##                            Comp.8      Comp.9      Comp.10      Comp.11
## Standard deviation     0.35702153 0.174687526 0.1031435150 0.0577993220
## Proportion of Variance 0.01158767 0.002774157 0.0009671441 0.0003037056
## Cumulative Proportion  0.99595499 0.998729150 0.9996962944 1.0000000000
```

Source: http://geog.uoregon.edu/bartlein/courses/geog495/lec16.html

(a) How many Principle Components are suitable to represent this dataset? [3]

(b) Based on the PCA outputs what can you conclude about the data? Discuss. [3]

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

(iv) Based on Scatter plot matrix in Fig. 1, what can you claim about the cities and related attributes? Discuss at least 3 findings. [6]

## Question 3 (20 marks)

(i) Which clustering technique would you recommend to cluster each of the datasets shown in Fig. 2? Justify your answer. [3 × 3]

(ii) Figure 3 shows a cluster dendogram of vehicles. What can you conclude from this cluster analysis? Discuss. [5]

(iii) Suppose PizzaHut gives you data of all home deliveries in Colombo (1 to 15) within the last 12 months. The dataset includes type of order, transaction value, order and delivery times, and delivery location (latitude and longitudes). To reduce the delivery cost and time, PizzaHut plans to establish 3 new locations in Colombo that will only handle home deliveries. Suggest a clustering-based solution to identify the best 3 locations to establish the new delivery centers. [6]

## Question 4 (20 marks)

Scatter plot matrix in Fig. 4 shows the relationship between a person's IQ score (PIQ) and his/her brain size, height, and weight.

```
lm(d$PIQ ~ d$Brain + d$Height + d$Weight)

Coefficients:
(Intercept)    d$Brain    d$Height    d$Weight
    111.4      2.06       -2.73       0.001
```

(i) By analyzing the scatter plots and regression model parameters what can you claim about the relationship between PIQ, Brain size, Weight, and Height? Justify your claims. [6]

(ii) Write the corresponding Multiple Linear Regression equation. [3]

(iii) Predict the PIQ of a person, where Brian = 81.69, Height = 64.5, and Weight = 118. [3]

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

(v) Would the application of a Nonlinear Regression Model enhance the accuracy of PIQ prediction? Discuss. [4]

## Question 5 (20 marks)

(i) Time series in Fig. 5 shows the monthly total retail sales in the USA. Note that data are in millions of dollars and not adjusted for inflation.

(a) Discuss what can be learned from the 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 retail sales for the next three months? Justify. [4]

(ii) Following is a title of an article appeared in DJBooth.net, a website for music lovers. [10]

*Music Sales Falling, Tour Revenue in Danger. Can the Industry Survive?*

Based on the data visualizations in Fig. 6 discuss whether there are sufficient statistics to back the above title.

---- END OF PAPER ---

# Annex – Figures

## Question 2

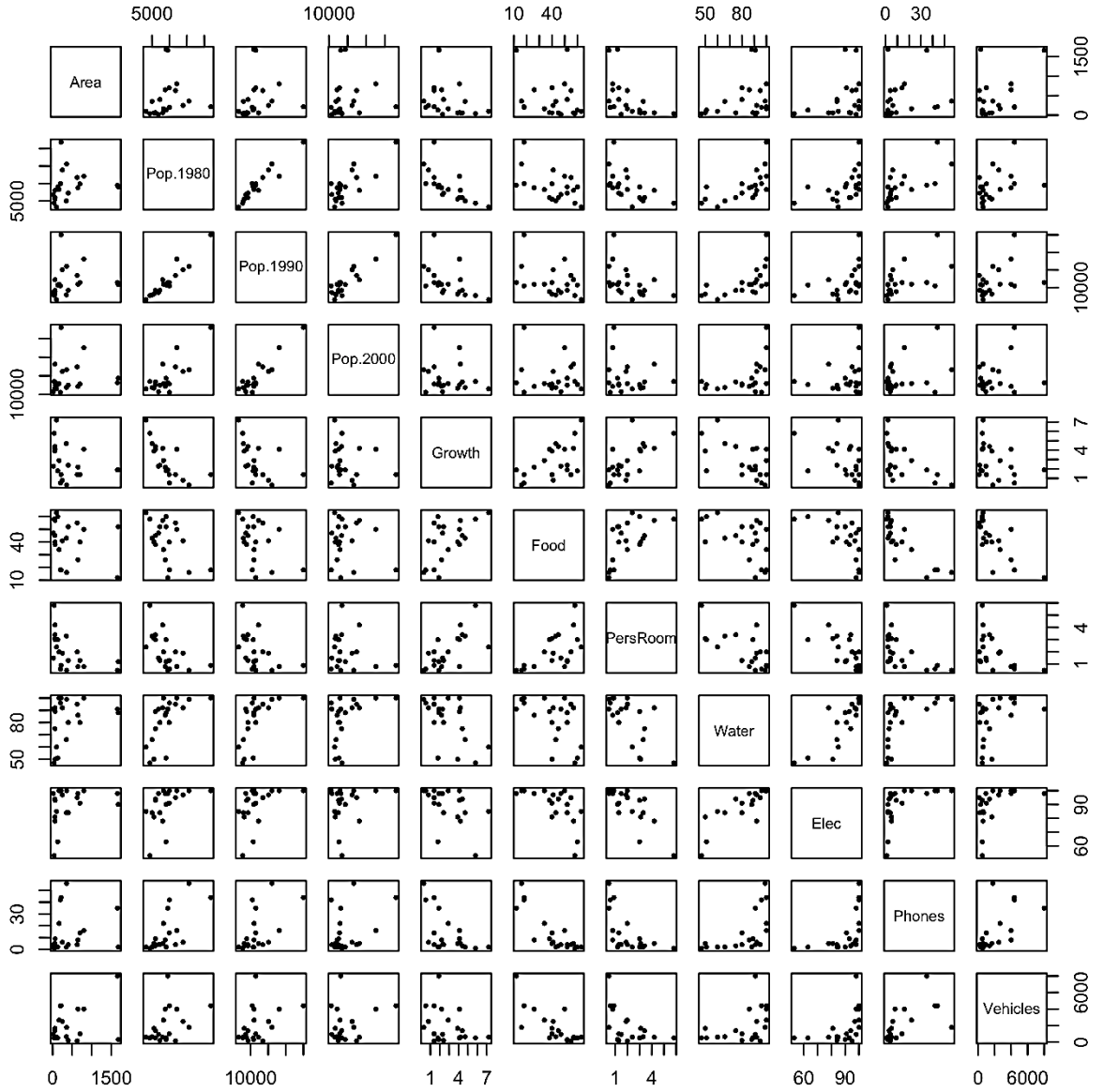


Figure 1

## Question 3

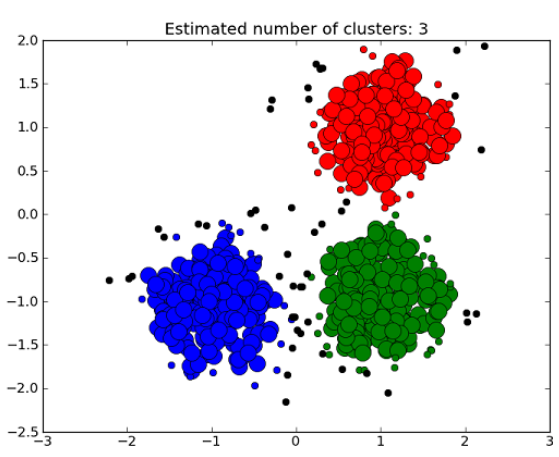


Figure 2(a)

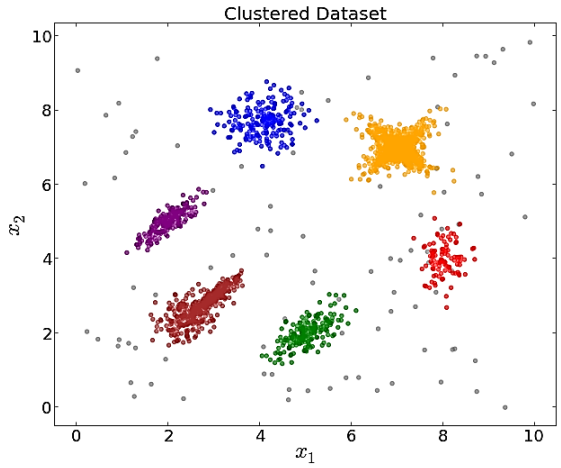


Figure 2(b)

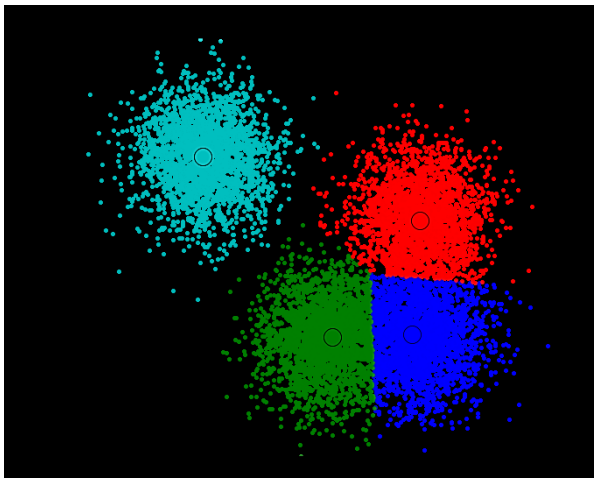

Figure 2(c)

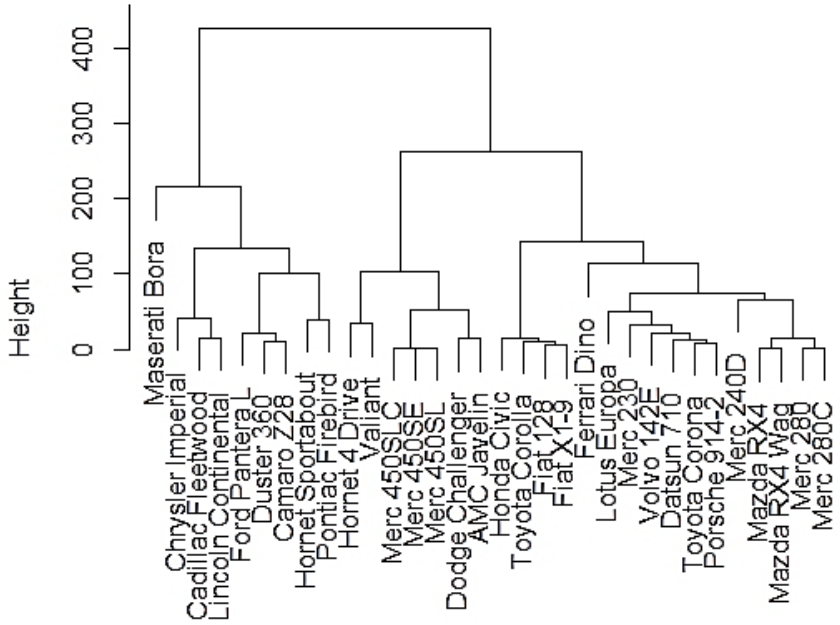


Figure 3 (Source: stackoverflow.com)

**Question 4**

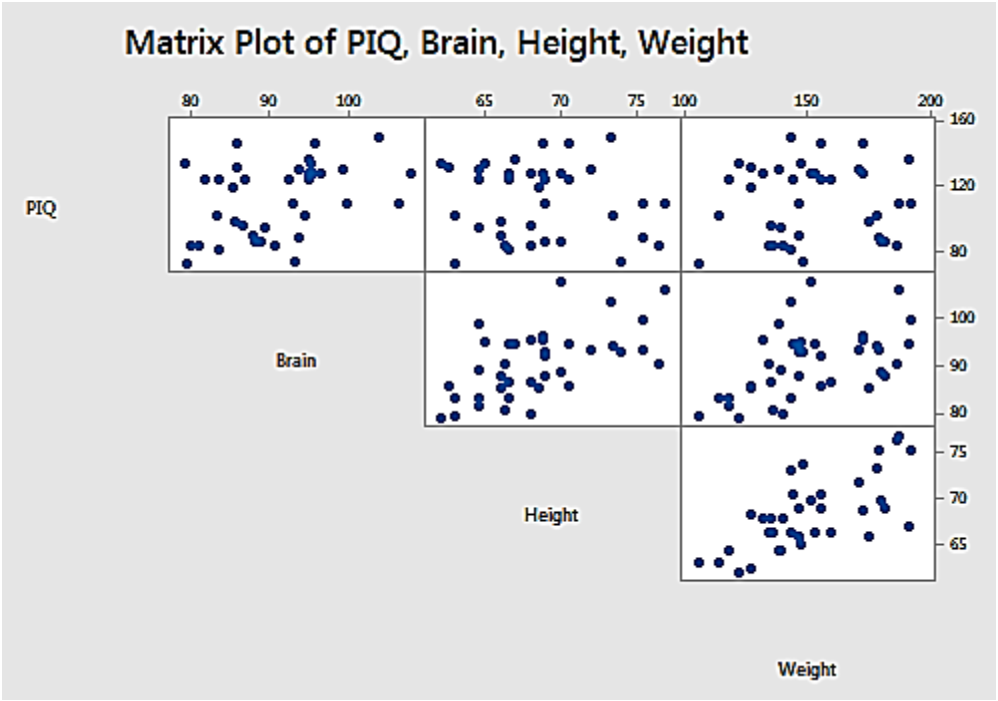


Figure 4 (Source: https://onlinecourses.science.psu.edu/stat501/node/331)

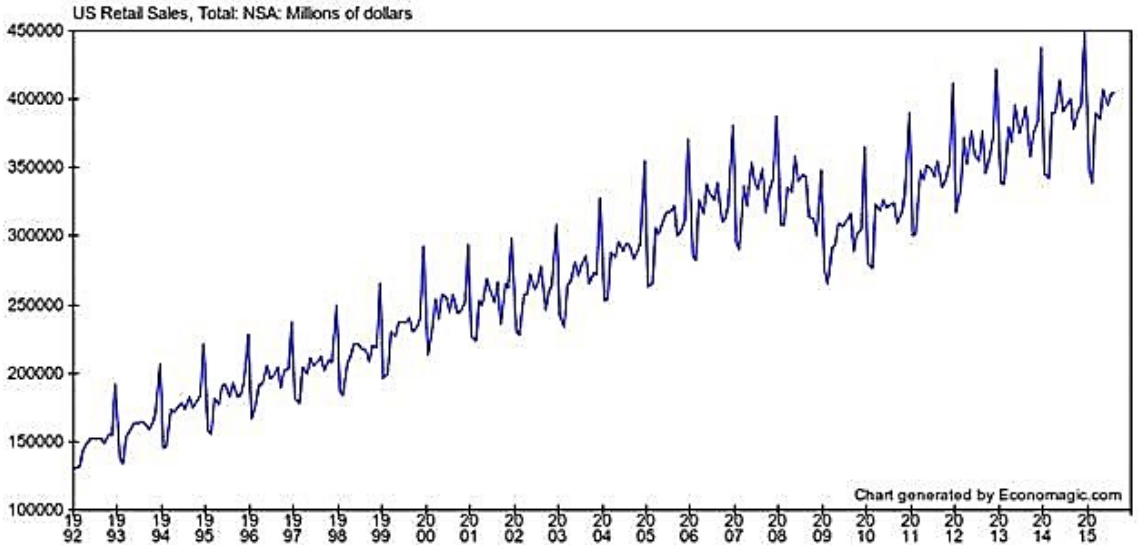


Figure 5

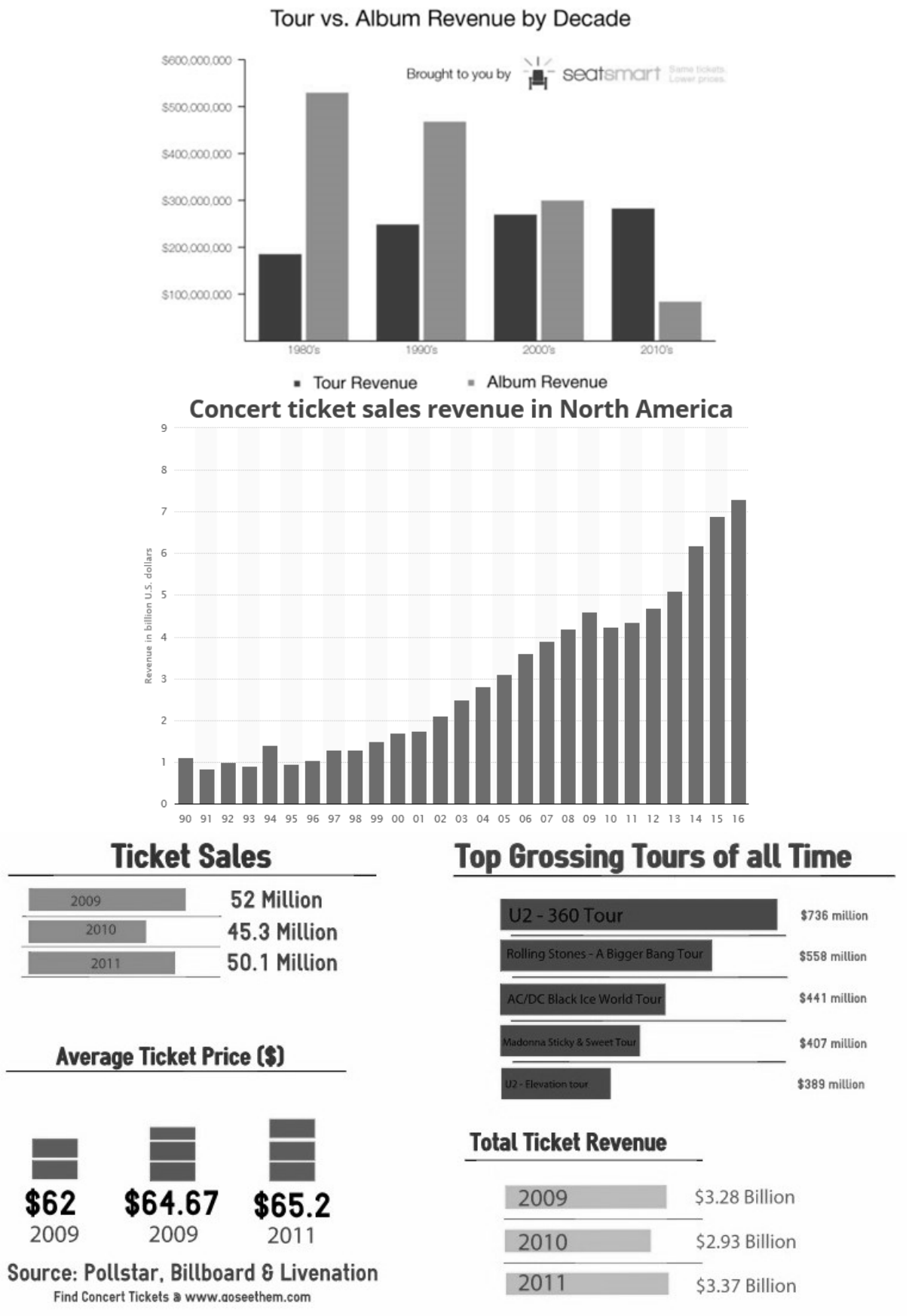
Tour vs. Album Revenue by Decade
Brought to you by seatsmart
$600,000,000
$500,000,000
$400,000,000
$300,000,000
$200,000,000
$100,000,000
1980's
1990's
2000's
2010's
Tour Revenue
Album Revenue
Concert ticket sales revenue in North America
Revenue in billion U.S. dollars
90 91 92 93 94 95 96 97 98 99 00 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16
Ticket Sales
2009
52 Million
2010
45.3 Million
2011
50.1 Million
Average Ticket Price ($)
$62
2009
$64.67
2009
$65.2
2011
Source: Pollstar, Billboard & Livenation
Find Concert Tickets @ www.goseethem.com
Top Grossing Tours of all Time
U2 - 360 Tour
$736 million
Rolling Stones - A Bigger Bang Tour
$558 million
AC/DC Black Ice World Tour
$441 million
Madonna Sticky & Sweet Tour
$407 million
U2 - Elevation tour
$389 million
Total Ticket Revenue
2009
$3.28 Billion
2010
$2.93 Billion
2011
$3.37 Billion

Figure 6