BUS240 Information Management **Fall, 2018**
Prof Carver

Visualization Assignment 2 – Tableau & ggplot2

Introduction: In this assignment, you and a partner will create a "faceted" visualizations of from one set of data twice: once in Tableau, and then again using R and ggplot2. Rather than use a template, you'll include a screenshot and your R code right in this document. Faceting refers to creating multiple small graphs in panels on one screen, where each panel focuses on one categorical variable.

The assignment is inspired by the "Economy" tab of the Tableau World Indicators workbook. For simplicity, we will not use any interactivity in the graphs for this exercise. On the Economy tab, you see GDP and GDP per capita for various countries. You can select the time range and display all of the nations for one region at a time. Each facet shows the change in the wealth measurements for a different country.

Phase 1 – Tableau (40 points)

1. Create a NEW TAB in the workbook that creates a similar faceted graph showing the aggregate trends in **Birth Rate**, **Life Expectancy, and Infant Mortality** for the 6 regions of the world, just for the years 2004 through 2010. Note that this must be ONE graph with multiple panels (not a dashboard or set of 3 separate pages). Give the tab a descriptive title, and use your judgment about the best plotting geometries, colors, and so forth.

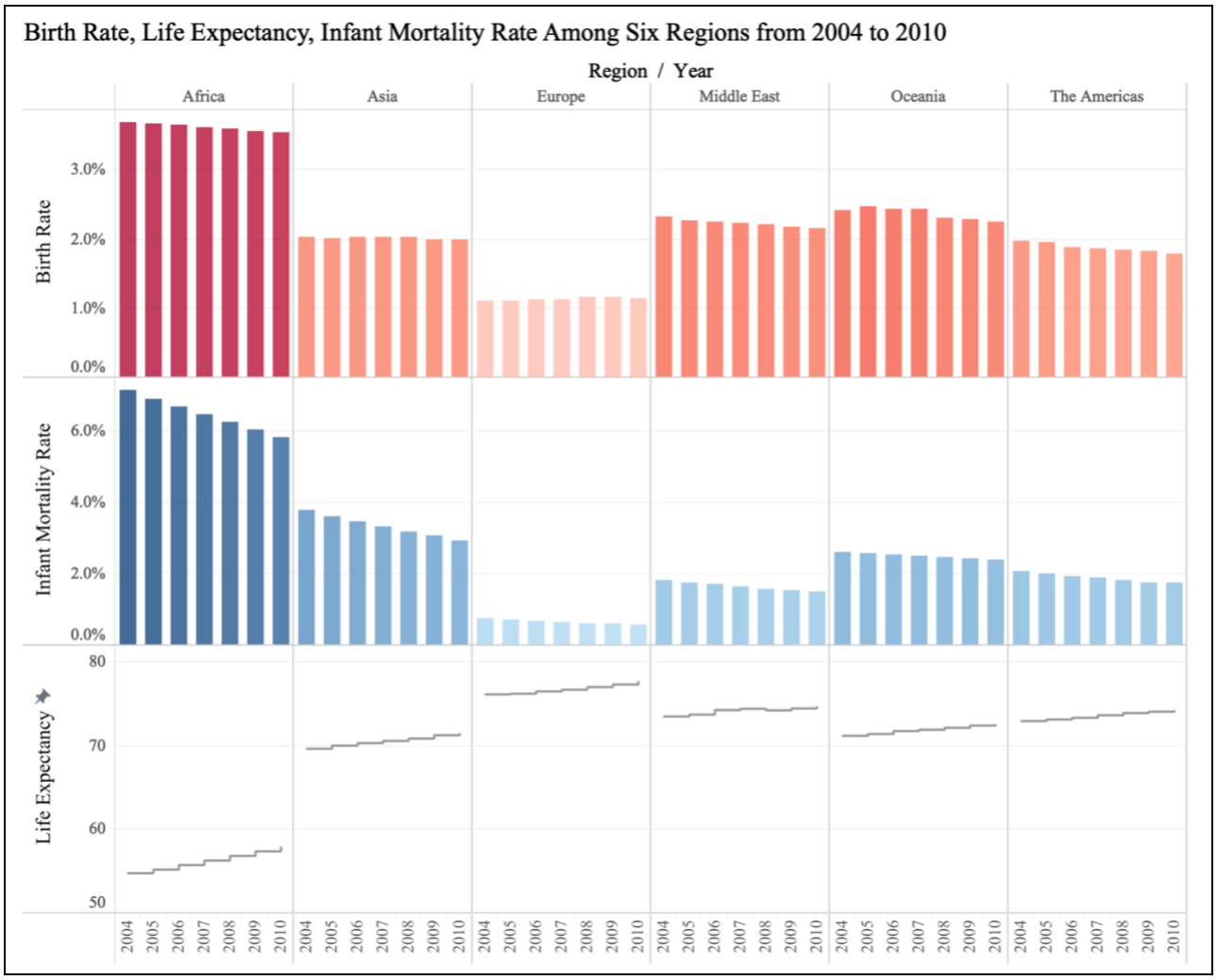
2. (15 pts) In a short paragraph, please explain what you both decided about the clearest way to create the plot. Write your response here (NOTE that the rectangle will enlarge as you type):

First and foremost, we used the data source from "World Indicators" to create the facets for regions to show the audience with or without any professional background the obvious overall comparison in the birth rate, infant mortality rate, and life expectancy among the six different regions of the world.

Next, we put the birth rate, infant mortality rate, and life expectancy on the rows and labeled them on the vertical axis, which is the best way to display these three continuous variables in our visualization analysis. First, we measured the birth rate by average, which means the average number of live births per 1000 specific population in a year. For example, on the graph, the birth rate of Africa in the year 2010 is 3.5%, which means that the average number of births in Africa is 35 births per 1000 population in 2010. Second, we measured the infant mortality rate by average, which means the average number of deaths of children under one year of age per 1000 live births in a particular population. For example, on the graph, the infant mortality rate of Africa in the year 2010 is 5.8%, which means that the average number of deaths of children under one year of age in Africa is 58 deaths per 1000 live births in 2010. Third, we measured the life expectancy by average, which means the average time a person is expected to live. Then, we put the years (filtered from 2004 to 2010) on the columns and labeled them on each facet to show the aggregate trends of those three continuous variables in a certain time range. Also, we set the values on the vertical axis of the life expectancy starting from 50 years and ending at 80 years because all the data of life expectancy from the World Indicators are in this range. By specifying the starting and ending values, we can find clearer trends in the life expectancy.

Further and more importantly, we used bar graphs as the geometric shapes of Birth Rate and Infant Mortality Rate, on which the length of the rectangles represents the values; and we plotted lines as the geometric shapes of the life expectancy. In this case, we can not only make the comparison for the three continuous variables in different years and regions but also explore the relationships among the birth rate, the infant mortality rate, and the life expectancy. For example, the graph shows that the birth rate (3.7% to 3.5%) and the infant mortality rate (7.1% to 5.8%) in Africa from 2004 to 2010 are both decreasing, and the life expectancy in Africa during the same period is increasing from 55 years old to 57 years old. This finding from the graph can made the audience think more about whether the decreasing birth rate and infant mortality rate are related to an increasing life expectancy. Also, we used the color legend for the values of birth rate and the infant mortality rate based on the brightness of the colors on the bar plots. We set that darker color reflects the higher value and lighter color shows the lower value, which is more effective to compare the values of the three continuous variables at first glance.

3. (10 pts) When the graph is complete, move into Presentation Mode, and then take a screen capture of the image that you have created. Paste that screen capture here:



4. (15 pts) Offer your insights and conclusions about birth rates, life expectancy, and infant mortality in the 6 regions of the world during the period in question. In other words, what story do you see in this visualization?

1. Birth Rates: The births rates in each region around the world are decreasing (Africa, Middle East, Oceania, and the Americas) or being relative stable (Asia and Europe) through the years. Among the six regions from 2004 to 2010, Africa has the highest overall birth rates but has an apparent decreasing trend (from 3.7% to 3.5%) through the years. The possible explanations for this result are that many children die at an early age due to diseases and famine, and many children are needed for agricultural activities in poor areas. As time changed, the government in Africa improved the medical care and diet so that fewer children are needed. On the contrary, Europe has the lowest overall birth rates (comes with stable rates around 1.1% and 1.2% through years) possibly because of the family planning and good healthcare services in the developed countries in Europe. Moreover, the status of women has been improved in a large scale so that there are more women prefer the later marriages.

2. Infant Mortality Rates: The infant mortality rates in the whole world are decreasing year over year. Among the six regions from 2004 to 2010, Africa has the highest overall infant mortality rates but experiencing the sharp decreases (from 7.1% to 5.8%, sharper than the reduction in the birth rates) through the years. The possible explanations for this result are starvation, infectious disease (AIDS) and lack of medical knowledge in poor areas. Through time changed, public health infrastructure, improvement in the water and food supply, and sanitation efforts from the African government helped reduce these forms of mortality. On the contrary, Europe has the lowest overall infant mortality rates and keeps decreasing through the years (from 0.8% to 0.6%) possibly because of the well-developed healthcare systems and reliable food supply.

3. Life expectancy: We found from the graph that the life expectancy for the whole world is increasing year over year. Among the six regions from 2004 to 2010, the life expectancy each year in Africa is the lowest but increased from 55 years to 57 years; and the life expectancy each year in Europe is the highest that grew from 73 years to 77 years.

After combing the findings and explanations we mentioned above, it is clear to conclude that the birth rates and infant mortality rates are both high in the poor and developing regions with inadequate medical care and famine, especially in Africa, and these leading causes of the high rates also reflect a lower life expectancy. On the contrary, the birth rates and infant mortality rates are both low in the wealthy regions with reliable food supply and developed health care, and these leading causes of low rates also reflect a higher life expectancy. Especially in Europe, the region is facing the aging population.

Phase 2 – R (40 points)

On LATTE, in the DATA tab, you will find a csv file containing the data from Tableau. Using R and package ggplot2 (along with other packages you find helpful), read in the data and create one visualization that is as similar as possible to the viz that you created in Tableau. The layout and information content should be nearly identical; do your best to match colors, line widths, backgrounds, and other visual features. Do not worry about matching the fonts.

1. (10 pts) Briefly explain the logic of your ggplot code chunk – in other words, what did you decide about aesthetics, geometries, and faceting.

First, we further processed the data on the base of the given R file. We divided `birthrate` and `infant` by 100 in order to show them in the percentage format in our plots. Then, we converted the modified grouped data frame, `wi3` to a typical data frame. In this way, we can better manipulate the rows and columns in it.

Overall, we mapped three graphs and merged them using the `vplayout` function. This function was defined to plot multiple graphs on a pre-divided layout. In this assignment, to make the visualization similar to the one we created using Tableau, we divided the layout into three rows. Similar to the Tableau visualization, we put the bar graph for the birth rate on the first row, put the bar graph for the infant mortality rate on the second row, and placed the line graph for the life expectancy on the bottom.

For each of the three graphs, we set x-axis on the year from 2004 to 2010 and y-axis on the three targeted variables. Using the `facet.wrap` function, we divided each graph into six segmentations based on six regions. Therefore, our final R plot looks similar to the Tableau visualization that we have three rows for the three factors and each row contains six sub-graphs indicated by regions.

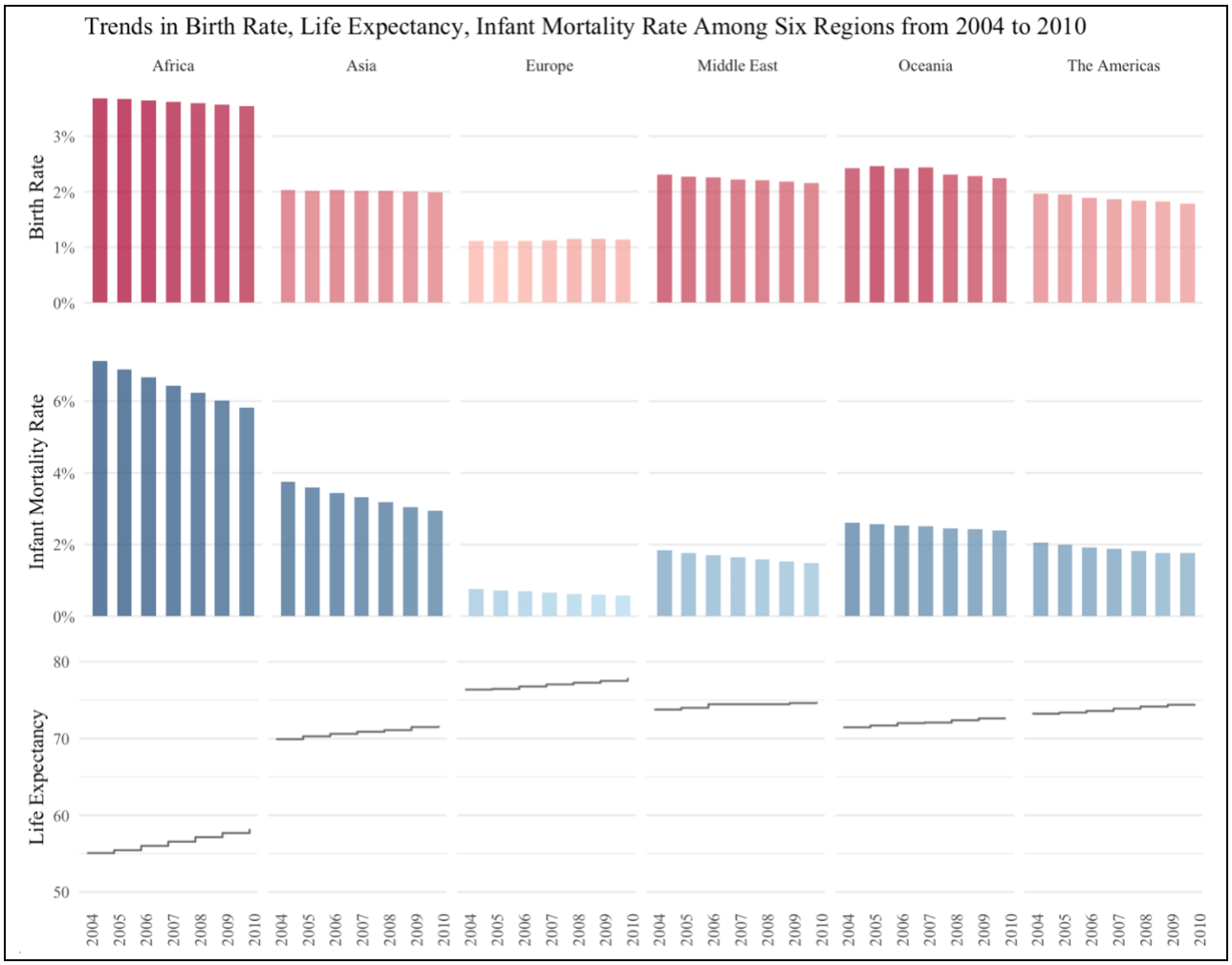
From the perspective of geometry layer, we followed the same decision as before that we used bar graphs to plot the birth rate and infant mortality rate on the year for different regions, and we used a line graph to make the plot for the life expectancy on the same scale.

From the perspective of faceting, we used the `facet.wrap` function to wrap a one dimension sequence of panels into two dimensions by adding `region` as the second dimension. This function helped us to imitate the Tableau manipulation of adding a second column. The result of faceting shows the data for each region in six separate sub-graphs.

As for the aesthetic layer, we improved it from the following seven aspects:

1. For the two bar graphs, we set the `fill` to the value of y (birth rate and infant mortality rate). In this way, we can use the shade of color to identify the value of data. To use gradient color to fill the bars, we created two new color palettes using the `colorRampPalette` function with the colors we extracted from Tableau. Hence, we successfully match the colors with Tableau Visualization.
2. Also, we set the width of bars to 0.6 and the transparency of bars to 0.8. These settings helped us to improve the similarity between the two graphs.
3. For the line graphs, we chose geom_step to simulate the stepped line graph in Tableau.
4. We added labels for x and y-axes then rotated the labels of years by 90 degrees.
5. We changed the font of the graphs to Times New Roman.
6. Moreover, the y-axes were rescaled based on the chosen scale for the Tableau visualizations.
7. Lastly, we applied theme_minimal to the graphs. This defaulted theme has a white background and simple grid lines, which is very similar to Tableau visualization. Also, by setting panel.grid.minor.y = element_blank(), we removed unnecessary grid lines.

2. (15 pts) Copy and paste your finished visualization here:



3. (15 pts) Copy and paste your R code here (use Paste > Keep text only):

```
# Divide bithrate and infant by 100 to show them in percentage format
wi3 <- wi3 %>% mutate(birthrate = birthrate/100, infant = infant/100)

# Convert wi3 into a data frame
wi3 <- as.data.frame(wi3)

# Create two new color palettes
red <- colorRampPalette(c("#FFBEB2", "#AE123A"))
blue <- colorRampPalette(c("#B9DDF1", "#2A5783"))

# The plot for birth rate
p1 <- wi3 %>% ggplot(aes(x = Yr, y = birthrate, fill = factor(birthrate))) +
  geom_bar(stat = "identity", alpha = 0.8, width = 0.6) +
  facet_wrap(.~Region, nrow = 1) +
  theme_minimal(base_family = "Times New Roman") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1), legend.position="none") +
```

```
  labs(x = NULL, y = "Birth Rate") +
  ggtitle("Trends in Birth Rate, Life Expectancy, Infant Mortality Rate Among Six Regions from 2004 to 2010") +
  scale_y_continuous(limits = c(0, 0.038), labels = percent_format(accuracy = 1)) +
  scale_fill_manual(values = red(41)) +
  scale_x_continuous(breaks = NULL) +
  theme(panel.grid.minor.y = element_blank())

# The plot for infant mortality rate
p2 <- wi3 %>% ggplot(aes(x = Yr, y = infant, fill = factor(infant))) +
  geom_bar(stat = "identity", alpha = 0.8, width = 0.6) +
  facet_wrap(.~Region, nrow = 1, strip.position = NULL) +
  theme_minimal(base_family = "Times New Roman") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1), legend.position="none", strip.text.x = element_blank()) +
  labs(x = NULL, y = "Infant Mortality Rate") +
  scale_y_continuous(limits = c(0, 0.075), labels = percent_format(accuracy = 1)) +
  scale_fill_manual(values = blue(42)) +
  scale_x_continuous(breaks = NULL) +
  theme(panel.grid.minor.y = element_blank())

# The plot for life expectancy
p3 <- wi3 %>% ggplot(aes(x = Yr, y = life_exp)) +
  geom_step(group = 1, color = "#686868") +
  facet_wrap(.~Region, nrow = 1, strip.position = NULL) +
  theme_minimal(base_family = "Times New Roman") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1), legend.position="none", strip.text.x = element_blank()) +
  labs(x = NULL, y = "Life Expectancy") +
  scale_y_continuous(limits = c(50, 80)) +
  scale_x_continuous(breaks = c(2004, 2005, 2006, 2007, 2008, 2009, 2010)) +
  theme(panel.grid.major.x = element_blank(), panel.grid.minor.x = element_blank())

# Define a function to divide the layout
vplayout <- function(x,y){viewport(layout.pos.row = x, layout.pos.col = y)}

# Create a new grid for the plots
grid.newpage()
pushViewport(viewport(layout = grid.layout(3,1))) #change the composition of the grid
print(p1, vp = vplayout(1,1))
print(p2, vp = vplayout(2,1))
print(p3, vp = vplayout(3,1))
```

Phase 3 – Comparisons (20 pts)

1. Write a short paragraph comparing Tableau and R as platforms for accomplishing this particular task. In your comments, think about ease of use, range of options available to you as the graph creator, ability to re-use the work you've done with a different set of data (reproducibility), and about human perception and cognition. Discuss the strengths and weakness of the two "tools" for this assignment.

Generally, Tableau is commonly acknowledged as a visualization tool, and R is widely known as a programming language for statistical computing and graphics. Therefore, intuitively, Tableau is more suitable for making gorgeous plots with simple statistical computation while R is appropriate for making succinct graphs with relatively more complicated calculations.

From the accomplishment of this particular task, we first made a plot in Tableau, then mapped a similar one in R. We figured out that, as the calculation required for this task is very simple, we can easily make a nice-looking visualization by dragging relative variables to the `rows` and `columns` section in Tableau, and Tableau can smartly assign an appropriate geometry graph to the variables. However, in R, we need to set the geometry layer manually by adding geom_bar or geom_step. Moreover, the aesthetic parameters can easily be adjusted by clicking defaulted buttons in Tableau. However, those parameters need to be coded manually and carefully in R because each of them should be concluded in different sentences. For example, when setting limitations on the y-axis, in Tableau, we can easily double click the labels on the left to set the fixed values; however, in R, we need to write a separate sentence (scale_y_continuous(limits = c(50, 80))) for this simple manipulation. Lastly, Tableau can create graphs with simple grid lines, but in R, we have to remove the unnecessary horizontal and vertical lines by coding (panel.grid.minor.y = element_blank()). Those minor grid lines often affect visual aesthetics.

According to the analyses above, we made the following conclusions on Tableau and R.

Tableau is an excellent option for pattern discovery using data visualization. From the perspective of beautification, using Tableau is a great choice because it has simple operation interface and abundant operation options. We can easily find the right pane to conduct the expected manipulations. Moreover, Tableau offers us the possibility of displaying the dynamic changes of data, which is impossible in R. However, when the raw data is complicated and need complex processing, Tableau is then not the first choice. Although it can do some necessary statistical computations, and it does have default quick-calculations, it cannot handle the nested data sources and web scrapings or do the complicated statistical analysis.

R is the right choice for data wrangling and manipulation. It can deal with massive amount of data in a second and conduct complex mathematical and statistical analysis. It is effortless for us to change the data type for each variable, and it is also convenient to create new variables based on the existing one. R enabled us to shrink the huge data frame `wi` into a four-column data frame `wi3`, which is clear for us to pick-up needed factors for the plot. However, when we moved to the visualization part after data cleaning, we spent double time in R than in Tableau. Instead of doing simple clicking in Tableau, we manipulated every aesthetic parameter by coding. Also, it is not easy to put three graphs together, so that we had to define an external grid function to realize the ideal manipulation.

In conclusion, for this assignment, we think Tableau is much easier to use for visualization.