Project2_C

true
true

3/10/2019

Introduction :-

As part of project 2, this is the third messy data which needs to cleaned and tranformed to be ready to do the analysis.

Problem Statment :-

This data set contains wine ratings of different varieties which are from different countries,the data has lot of blank values for wine points/country which needs to be cleaned and the data is mismanaged which needs to first sorted and arranged in way, before any meaningfull analysis can be done.

Solution :-

We are using below libraries in our quest to resolve the above problem:- *knitr*
tidyr
dplyr
kableExtra
ggplot2

Steps performed on cleaning, tranforming and doing Analysis on the data

1) Reading the data from local directory using read.csv function.
2) Setting the blank data to NA and then we selected the relevant columns for our cleaning up and analysis using select function from tidy.

```
raw_df <- read.csv('winemag-data-130k-v2.csv', stringsAsFactors=F)
raw_df[raw_df==""] <- NA

wine_data <- select(raw_df , 2,5,6,7,13,14)
kable(head(wine_data)) %>%
  kable_styling(bootstrap_options = c("striped","hover","condensed","responsive"),full_width   = F,posi
  row_spec(0, background ="gray")
```

country	points	price	province	variety	winery
Italy	87	NA	Sicily & Sardinia	White Blend	Nicosia
Portugal	87	15	Douro	Portuguese Red	Quinta dos Avidagos
US	87	14	Oregon	Pinot Gris	Rainstorm
US	87	13	Michigan	Riesling	St. Julian
US	87	65	Oregon	Pinot Noir	Sweet Cheeks
Spain	87	15	Northern Spain	Tempranillo-Merlot	Tandem

3) Then we removed the records which had Country as NA, using drop_na from tidy library.
4) Next step we are calculating the average points per country wise , using group_by, summarize and applying mean on points for that country all these functions are from dplyr library.
5) we used glinmpse function from tidyr to have a snapshot of our final data set.

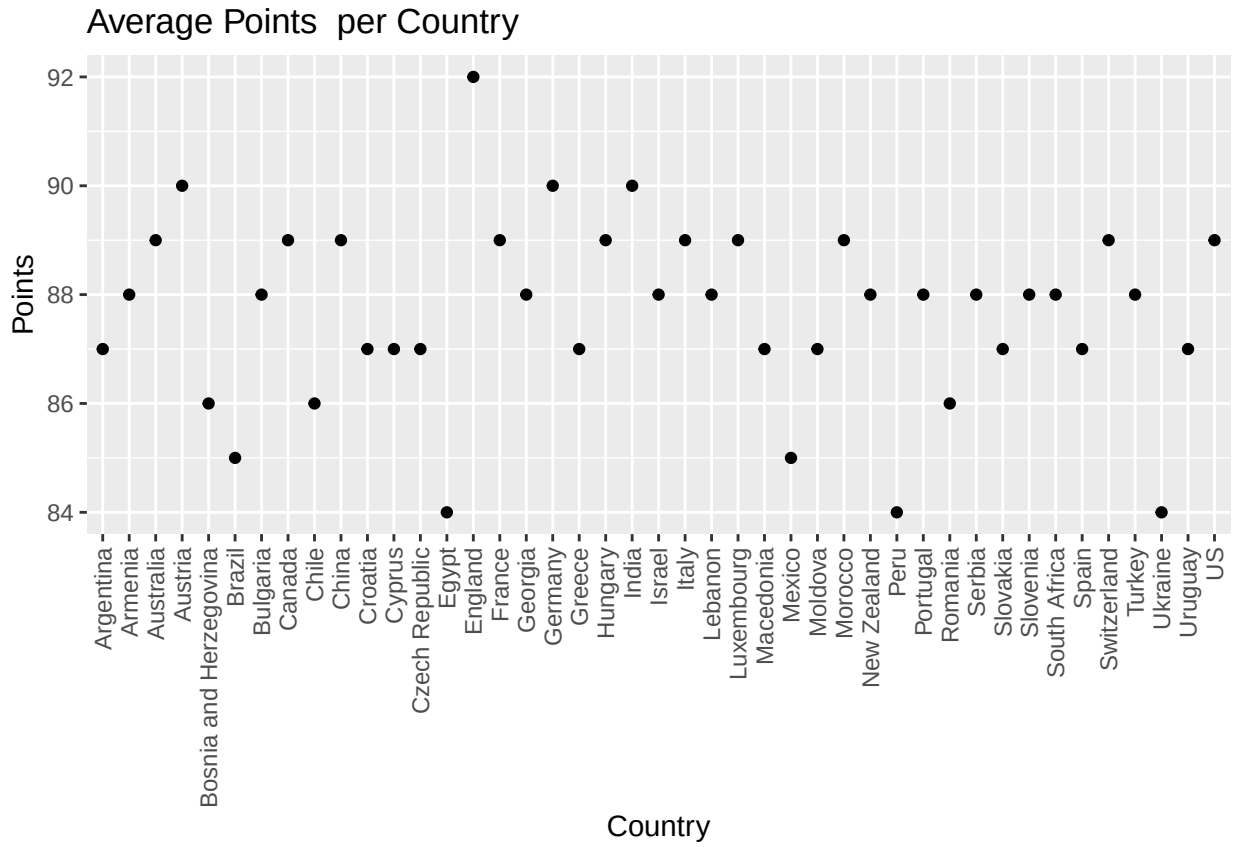
```
wine_data <- wine_data %>%   drop_na(country)
wine_data_variety <- wine_data %>%
  group_by(country ) %>%
  summarise(avg_points = round(mean(points)))

kable(head(wine_data_variety)) %>%
  kable_styling(bootstrap_options = c("striped","hover","condensed","responsive"),full_width    = F,posi
  row_spec(0, background ="gray")
```

country	avg_points
Argentina	87
Armenia	88
Australia	89
Austria	90
Bosnia and Herzegovina	86
Brazil	85

6) Using the ggplot2 library we plot the average points against Country names to do our analysis.

```
wine_data_variety %>%
  ggplot(aes(x=country, y=avg_points)) +
  geom_point() + geom_line() +
  labs(title = "Average Points  per Country", x="Country", y="Points") +
  theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5))
```

```
## geom_path: Each group consists of only one observation. Do you need to
## adjust the group aesthetic?
```



Summary

Looking at the plotted graph we can say that the wines from England has been adjudged the best with 92 average wine tasting points, closely followed by Austria, Germany and India with 90 points. whereas Egypt,Peru and Ukraine wines have the least wine tasting average points.