Review, and Payment Type

Team 1

11/25/2020

## Questions answering

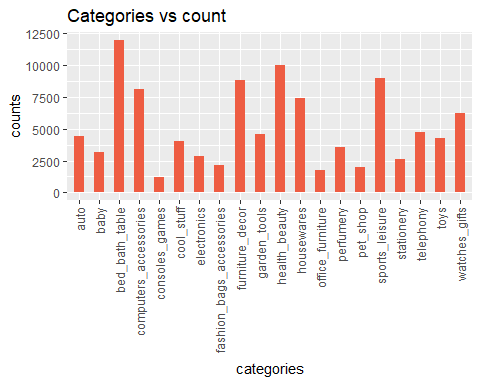
##Q2. What categories have the most order (based on product\_id)? housewares 7380  
computers\_accessories 8151  
furniture\_decor 8833  
sports\_leisure 9005  
health\_beauty 10030  
bed\_bath\_table 11990

As shown, bed bath table is the most order product category.

#products categories based on product\_ids   
  
try\_1 <- olist %>%   
 group\_by(product\_category\_name\_english)%>%  
 summarise(result=(product\_id))

## `summarise()` regrouping output by 'product\_category\_name\_english' (override with `.groups` argument)

#Most selling categories top 20  
x=as.data.frame(count(try\_1))  
x\_1 <- x[order(x$n),][71:52,]  
  
#Bar chart most selling categories top 20  
ggplot(x\_1, aes (x=product\_category\_name\_english, y=n))+  
 geom\_bar(stat= "identity", width=0.5 , fill="tomato2")+  
 labs(x= "categories", y="counts", titles =paste ("Categories vs count"))+theme(axis.text.x=element\_text(angle=90,hjust=1,vjust=0.5))



##Q3. What categories have the highest sum reviews? These are the top 10 categories of the highest sum review scores:

bed\_bath\_table 46366  
health\_beauty 41315  
sports\_leisure 36856  
furniture\_decor 34407  
computers\_accessories 32002  
housewares 29854  
watches\_gifts 24856  
telephony 18612  
garden\_tools 18412  
auto 17770

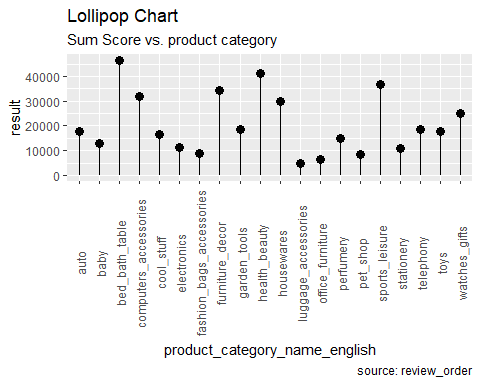
try\_2 <- olist %>%   
 group\_by(product\_category\_name\_english)%>%  
 summarise(result=sum(review\_score))

## `summarise()` ungrouping output (override with `.groups` argument)

x\_3 <- try\_2[order(try\_2$result),][71:52,]  
x\_3

## # A tibble: 20 x 2  
## product\_category\_name\_english result  
## <chr> <dbl>  
## 1 bed\_bath\_table 46366  
## 2 health\_beauty 41315  
## 3 sports\_leisure 36856  
## 4 furniture\_decor 34407  
## 5 computers\_accessories 32002  
## 6 housewares 29854  
## 7 watches\_gifts 24856  
## 8 telephony 18612  
## 9 garden\_tools 18412  
## 10 auto 17770  
## 11 toys 17734  
## 12 cool\_stuff 16499  
## 13 perfumery 14797  
## 14 baby 12802  
## 15 electronics 11466  
## 16 stationery 10963  
## 17 fashion\_bags\_accessories 8933  
## 18 pet\_shop 8503  
## 19 office\_furniture 6288  
## 20 luggage\_accessories 4990

ggplot(x\_3, aes(x=product\_category\_name\_english, y=result)) +   
 geom\_point(size=3) +   
 geom\_segment(aes(x=product\_category\_name\_english,   
 xend=product\_category\_name\_english,   
 y=0,   
 yend=result)) +   
 labs(title="Lollipop Chart",   
 subtitle="Sum Score vs. product category ",   
 caption="source: review\_order") +   
 theme(axis.text.x = element\_text(angle=90, vjust=0.6))



##Q4. Who are the most reviewed sellers (based on review counts)?

Most reviewed sellers’ ID and number counts:

4a3ca9315b744ce9f8e9374361493884 335  
7c67e1448b00f6e969d365cea6b010ab 333  
1f50f920176fa81dab994f9023523100 296  
6560211a19b47992c3666cc44a7e94c0 287  
1025f0e2d44d7041d6cf58b6550e0bfa 240  
cc419e0650a3c5ba77189a1882b7556a 236  
da8622b14eb17ae2831f4ac5b9dab84a 186  
955fee9216a65b617aa5c0531780ce60 160 ea8482cd71df3c1969d7b9473ff13abc 158  
cca3071e3e9bb7d12640c9fbe2301306 139

try\_55 <- olist %>%   
 group\_by(seller\_id)%>%  
 summarise(result=count(review\_score))

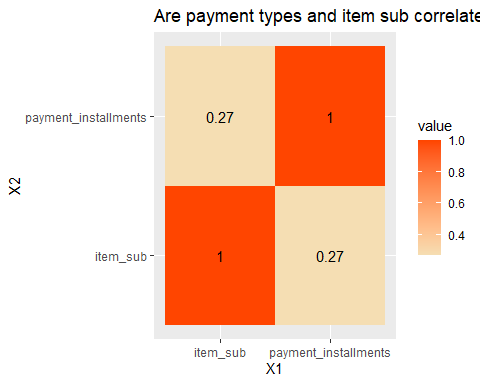
## `summarise()` ungrouping output (override with `.groups` argument)

x\_4 <- try\_55[order(try\_55$result),][3033:3014,]  
x\_4

## # A tibble: 20 x 2  
## seller\_id result  
## <chr> <int>  
## 1 4a3ca9315b744ce9f8e9374361493884 335  
## 2 7c67e1448b00f6e969d365cea6b010ab 333  
## 3 1f50f920176fa81dab994f9023523100 296  
## 4 6560211a19b47992c3666cc44a7e94c0 287  
## 5 1025f0e2d44d7041d6cf58b6550e0bfa 240  
## 6 cc419e0650a3c5ba77189a1882b7556a 236  
## 7 da8622b14eb17ae2831f4ac5b9dab84a 186  
## 8 955fee9216a65b617aa5c0531780ce60 160  
## 9 ea8482cd71df3c1969d7b9473ff13abc 158  
## 10 cca3071e3e9bb7d12640c9fbe2301306 139  
## 11 4869f7a5dfa277a7dca6462dcf3b52b2 139  
## 12 8b321bb669392f5163d04c59e235e066 136  
## 13 d2374cbcbb3ca4ab1086534108cc3ab7 135  
## 14 1835b56ce799e6a4dc4eddc053f04066 129  
## 15 3d871de0142ce09b7081e2b9d1733cb1 113  
## 16 897060da8b9a21f655304d50fd935913 111  
## 17 7a67c85e85bb2ce8582c35f2203ad736 103  
## 18 620c87c171fb2a6dd6e8bb4dec959fc6 102  
## 19 1900267e848ceeba8fa32d80c1a5f5a8 94  
## 20 88460e8ebdecbfecb5f9601833981930 90

##Q5. What is the relationship between payment value and installments? There is a positive 0.27 correlation between payment value and payment installments, a weak relationship between payment value and payment installments.

# payment installments and payment value correlation  
olist <- olist %>% mutate(item\_sub= order\_item\_id \* price)  
olist$item\_sub <- as.numeric(olist$item\_sub)  
cor.mat <- round(cor(olist[c("payment\_installments","item\_sub")]),2)   
  
melted.cor.mat <- melt(cor.mat)   
ggplot(melted.cor.mat, aes(x = X1, y = X2, fill = value)) +   
 scale\_fill\_gradient(low="wheat", high="orangered") +  
 geom\_tile() +   
 geom\_text(aes(x = X1, y = X2, label = value)) +  
 ggtitle("Are payment types and item sub correlated?")

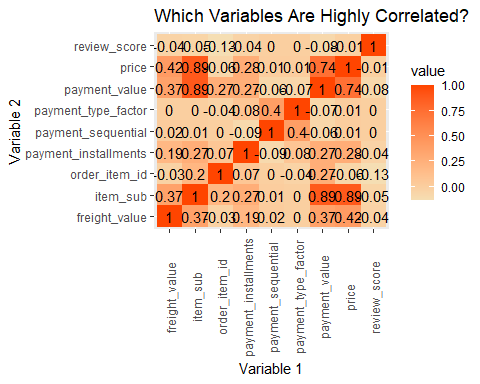


##Q6. Correlations between Payment types and other variables. We used a correlation heatmap to plot out the srongest positive and negative correlations. As indicated, “price” and “payment value” have the strongest correlation, with a positive 0.89. “review\_item\_id” and “review\_score” have the lowest correlation, with a negative 0.13.

#average payment with installment and without installment  
olist\_x<- olist[,c(12,15,16,17,18,19,20,22,35)]  
  
#olist\_x$review\_score <- factor(olist\_x$review\_score)  
#olist\_x$payment\_type <- factor(olist\_x$payment\_type)  
str(olist\_x)

## 'data.frame': 116581 obs. of 9 variables:  
## $ order\_item\_id : num 1 1 2 1 1 1 1 1 1 1 ...  
## $ price : num 125 113 113 125 107 ...  
## $ freight\_value : num 21.9 24.9 24.9 15.6 30.6 ...  
## $ payment\_sequential : num 1 1 1 1 1 1 1 1 1 1 ...  
## $ payment\_type : chr "credit\_card" "credit\_card" "credit\_card" "credit\_card" ...  
## $ payment\_installments: num 2 1 1 7 10 5 1 2 3 1 ...  
## $ payment\_value : num 147 276 276 141 138 ...  
## $ review\_score : num 4 1 1 3 4 4 4 2 3 2 ...  
## $ item\_sub : num 125 113 226 125 107 ...

olist\_x[olist\_x$payment\_type == 'boleto', 'payment\_type\_factor']<-0  
olist\_x[olist\_x$payment\_type == 'credit\_card', 'payment\_type\_factor']<-1  
olist\_x[olist\_x$payment\_type == 'debit\_card', 'payment\_type\_factor']<-2  
olist\_x[olist\_x$payment\_type == 'voucher', 'payment\_type\_factor']<-3  
olist$payment\_type\_factor <- factor(olist\_x$payment\_type\_factor)  
  
cor.mat\_1 <- round(cor(olist\_x[,-c(5)]), 2)  
  
melted.cor.mat\_1 <- melt(cor.mat\_1, varnames = c("Var1", "Var2"))  
  
ggplot(melted.cor.mat\_1, aes(x = Var1, y = Var2, fill = value)) +  
 scale\_fill\_gradient(low="wheat", high="orangered") +  
 geom\_tile() +   
 geom\_text(aes(x = Var1, y = Var2, label = value)) +  
 ggtitle("Which Variables Are Highly Correlated?") + xlab("Variable 1") + ylab("Variable 2")+  
 theme(axis.text.x = element\_text(angle=90, vjust=0.6))

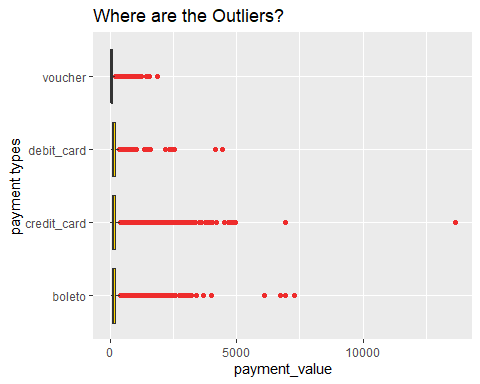


#Q7. What is the main payment type? Do customers who use a certain payment spend more than other customers? Credit card is the main payment type with a total of $15,481,976.1 from 2016 to 2018 Boleto is the second most used payment type with a total of $4,024,296.2  
Voucher is the third most used payment type with a total of $401,700.6  
Debit\_card is the least used payment type with a total of $250,077.9

Payment type and average payment value:

boleto 177.34427  
credit\_card 179.99786  
debit\_card 150.10676  
voucher 64.67567

# payment methods distributions and outliers  
ggplot(olist) +  
 geom\_boxplot(aes(x = payment\_type, y = payment\_value),   
 fill = "gold1", outlier.color = "firebrick2") +   
 coord\_flip() +   
 xlab("payment types") + ggtitle("Where are the Outliers?")



amount <-as.data.frame(olist %>%   
 group\_by(payment\_type)%>%  
 summarise(result=sum(payment\_value)))

## `summarise()` ungrouping output (override with `.groups` argument)

amount

## payment\_type result  
## 1 boleto 4024296.2  
## 2 credit\_card 15481976.1  
## 3 debit\_card 250077.9  
## 4 voucher 401700.6

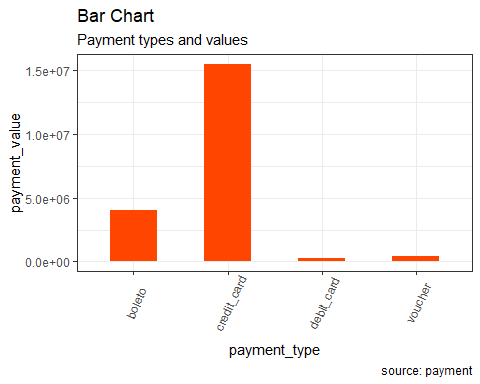
amount\_mean <-as.data.frame(olist %>%   
 group\_by(payment\_type)%>%  
 summarise(result=mean(payment\_value)))

## `summarise()` ungrouping output (override with `.groups` argument)

colnames(amount\_mean)[2] = "average amount"  
amount\_mean

## payment\_type average amount  
## 1 boleto 177.34427  
## 2 credit\_card 179.99786  
## 3 debit\_card 150.10676  
## 4 voucher 64.67567

# Draw plot  
theme\_set(theme\_bw())  
ggplot(olist, aes(x=payment\_type, y=payment\_value)) +   
 geom\_bar(stat="identity", width=.5, fill="orangered") +   
 labs(title="Bar Chart",   
 subtitle="Payment types and values",   
 caption="source: payment") +   
 theme(axis.text.x = element\_text(angle=65, vjust=0.6))

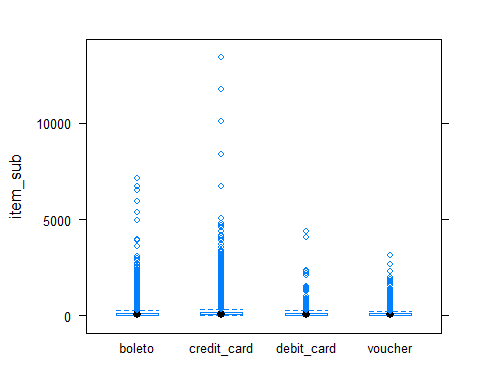


##Q8.Is there evidence that people are spending more when paying with credit card? Since the contribution of item sub from 4 payment types are right skewed and with several outliers, we decided to log item sub for further evaluation. From the ANOVA test, reject the null, there are statistically significance between variables. We went further for the Dunnet to see which payment type is different from the reference group. Result came out with Voucher is statistically significantly different from the reference group Boleto and credit card is also statistically significantly differernt form the reference group Boleto.

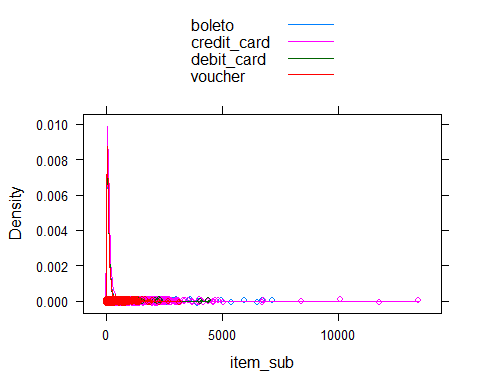
favstats( item\_sub~payment\_type, data=olist)

## payment\_type min Q1 median Q3 max mean sd n  
## 1 boleto 0.85 39.900 75.0 138.0000 7160 125.8332 213.4558 22692  
## 2 credit\_card 0.85 47.900 89.0 153.0000 13440 141.6228 220.8421 86012  
## 3 debit\_card 5.99 36.225 69.9 127.4425 4400 118.0174 229.2523 1666  
## 4 voucher 2.20 39.000 69.9 120.0000 3124 114.0773 164.2140 6211  
## missing  
## 1 0  
## 2 0  
## 3 0  
## 4 0

bwplot(item\_sub~payment\_type, data=olist)



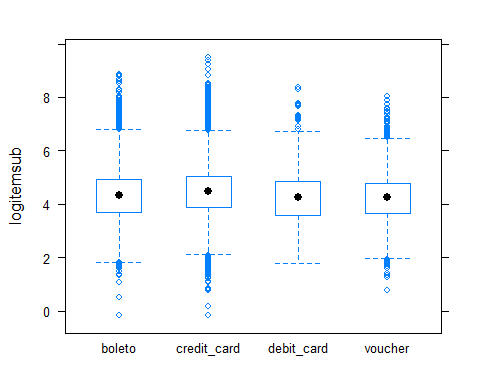
densityplot(~item\_sub, groups=payment\_type, auto.key = TRUE, data=olist)



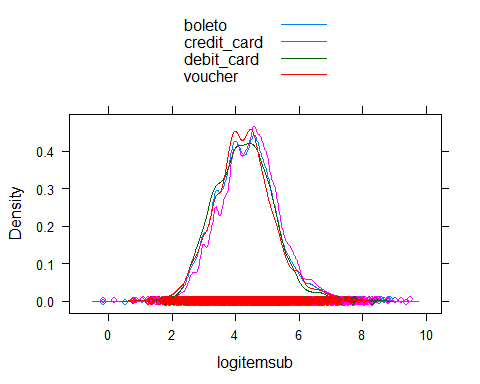
#log  
olist\_2= transform(olist, logitemsub= log(item\_sub))  
favstats(logitemsub~payment\_type, data=olist\_2)

## payment\_type min Q1 median Q3 max mean  
## 1 boleto -0.1625189 3.686376 4.317488 4.927254 8.876265 4.328411  
## 2 credit\_card -0.1625189 3.869116 4.488636 5.030438 9.505991 4.467361  
## 3 debit\_card 1.7900914 3.589692 4.247066 4.847665 8.389360 4.259022  
## 4 voucher 0.7884574 3.663562 4.247066 4.787492 8.046870 4.256076  
## sd n missing  
## 1 0.9429900 22692 0  
## 2 0.9365910 86012 0  
## 3 0.9191349 1666 0  
## 4 0.9344434 6211 0

bwplot(logitemsub~payment\_type, data=olist\_2)



densityplot(~logitemsub, groups = payment\_type, auto.key = TRUE, data=olist\_2)



#Significance  
ov1 <- anova(lm(logitemsub~payment\_type, data= olist\_2))  
ov1 #Reject null, at least one mean is different from others.

## Analysis of Variance Table  
##   
## Response: logitemsub  
## Df Sum Sq Mean Sq F value Pr(>F)   
## payment\_type 3 589 196.347 223.41 < 2.2e-16 \*\*\*  
## Residuals 116577 102456 0.879   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

summary(lm(logitemsub~payment\_type, data= olist\_2))

##   
## Call:  
## lm(formula = logitemsub ~ payment\_type, data = olist\_2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.6299 -0.6174 0.0100 0.5754 5.0386   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.328411 0.006223 695.510 < 2e-16 \*\*\*  
## payment\_typecredit\_card 0.138950 0.006996 19.860 < 2e-16 \*\*\*  
## payment\_typedebit\_card -0.069389 0.023796 -2.916 0.00355 \*\*   
## payment\_typevoucher -0.072335 0.013425 -5.388 7.13e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9375 on 116577 degrees of freedom  
## Multiple R-squared: 0.005716, Adjusted R-squared: 0.005691   
## F-statistic: 223.4 on 3 and 116577 DF, p-value: < 2.2e-16

#Dunnet prodcedure   
olist\_2$payment\_type <- factor(olist\_2$payment\_type)  
fit <- aov(logitemsub~payment\_type, olist\_2)  
set.seed(42)  
Dunnet <- glht(fit,linfct=mcp(payment\_type="Dunnett"))  
summary(Dunnet)

##   
## Simultaneous Tests for General Linear Hypotheses  
##   
## Multiple Comparisons of Means: Dunnett Contrasts  
##   
##   
## Fit: aov(formula = logitemsub ~ payment\_type, data = olist\_2)  
##   
## Linear Hypotheses:  
## Estimate Std. Error t value Pr(>|t|)   
## credit\_card - boleto == 0 0.138950 0.006996 19.860 <0.001 \*\*\*  
## debit\_card - boleto == 0 -0.069389 0.023796 -2.916 0.0105 \*   
## voucher - boleto == 0 -0.072335 0.013425 -5.388 <0.001 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## (Adjusted p values reported -- single-step method)

#Credit card, and Voucher are significantly different from rest of the payment types.