Question 4

```
pacman::p_load(tidyverse, broom, modelr, GGally, olsrr, pander, stargazer)
```

```
print(knitr::opts_knit$get("rmarkdown.pandoc.to"))
```

```
## [1] "latex"
```

```
hertz_data <-
  read_csv(file.path(data_path, "cust_survey_transaction.csv")) %>%
  rename(`Total_charge_USD` = `Total _charge_USD`)
```

```
hertz_data %>%
  select_if(function(x) any(is.na(x))) %>%
  summarise_all(funs(sum(is.na(.)))) %>%
  gather(Column, `Missing Values`) %>%
  pander(justify = c('left', 'right'))
```

Column	Missing Values
Overall_Exper	28,045
Staff_Courtesy	2,260
Speed_of_Service	2,264
Veh_Equip_Condition	2,265
Trans_Billing_as_Expected	2,281
Value_for_the_Money	2,283
rent_loc_type	2,289
rent_loc_name	20
cust_tier_code	251

We will not use the `Overall_Exper` column because 28,045 records have no response - 35% of the customer responses.

A

We treat the survey questions continuous variables, though we know they are actually ordinal and discrete.

```
base_formula <- as.formula(Recom_mend_Hertz ~ Staff_Courtesy +
                           Speed_of_Service + Trans_Billing_as_Expected +
                           Value_for_the_Money + Total_charge_USD +
                           Veh_Equip_Condition + Survey_checkout_diff)
lm1 <- lm(base_formula, data = hertz_data)
regression_results(lm1, title = "Full Model Results")
```

We see that increases in all survey questions and an increase in `Total_charge_USD` are associated with an increase in response to recommending Hertz. We note that `Survey_checkout_diff` is significant at the 95% but not the 99% confidence. For a more parsimonious model, we remove this variable.

Multicolinearity

We would expect that many survey questions are correlated thus resulting in issues with multicolinearity. One measure of multicolinearity is *variance inflation factors* (VIF) - a measure of how much the variance of

Table 2: Full Model Results

	Dependent variable:
	Recom_mend_Hertz
Constant	-0.576^{***}
	p = 0.000
Staff_Courtesy	0.239^{***}
	p = 0.000
Speed_of_Service	0.198^{***}
	p = 0.000
Trans_Billing_as_Expected	0.161^{***}
	p = 0.000
Value_for_the_Money	0.294^{***}
	p = 0.000
Total_charge_USD	0.0001^{***}
	p = 0.000
Veh_Equip_Condition	0.179^{***}
	p = 0.000
Survey_checkout_diff	-0.003^{**}
	p = 0.023
Observations	78,440
R^2	0.671
Adjusted R^2	0.671
Note:	*p<0.1; **p<0.05; ***p<0.01

each regression coefficient β_k is inflated by the existence of correlation among the predictor variables in the model[1]. There are lots of "rules-of-thumb" about what qualifies as a VIF that is indicates multicollinearity. A VIF of 4 is often indicates a need to investigate and while that's not the case here, can should still investigate further.

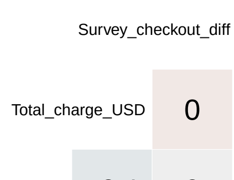
$$VIF = \frac{1}{1 - R_k^2}$$

```r
ols_vif_tol(lm1) %>% pander()
```

Variables	Tolerance	VIF
Staff_Courtesy	0.4972	2.011
Speed_of_Service	0.5193	1.926
Trans_Billing_as_Expected	0.5803	1.723
Value_for_the_Money	0.4811	2.079
Total_charge_USD	0.9892	1.011
Veh_Equip_Condition	0.6633	1.508
Survey_checkout_diff	0.9949	1.005

We then look at a correlation plot of each of the variables (including `Overall_Exper`) in the model:

```r
hertz_data %>%
  select(Recom_mend_Hertz, Overall_Exper, Staff_Courtesy, Speed_of_Service,
         Veh_Equip_Condition, Trans_Billing_as_Expected,  Value_for_the_Money,
         Total_charge_USD, Survey_checkout_diff) %>%
  ggcorr(label = TRUE, size = 2, hjust = 0.75, layout.exp = 1)
```



We then perform step-wise backwards elimination (using p-values) to remove variables from the model:

```
lm2 <- update(lm1, . ~ . - Survey_checkout_diff)
regression_results(lm1, lm2, title = "Model Comparison")
```

Table 4: Model Comparison

	Dependent variable:	
	Recom_mend_Hertz	
	(1)	(2)
Constant	-0.576^{***}	-0.587^{***}
	p = 0.000	p = 0.000
Staff_Courtesy	0.239^{***}	0.239^{***}
	p = 0.000	p = 0.000
Speed_of_Service	0.198^{***}	0.198^{***}
	p = 0.000	p = 0.000
Trans_Billing_as_Expected	0.161^{***}	0.161^{***}
	p = 0.000	p = 0.000
Value_for_the_Money	0.294^{***}	0.294^{***}
	p = 0.000	p = 0.000
Total_charge_USD	0.0001^{***}	0.0001^{***}
	p = 0.000	p = 0.000
Veh_Equip_Condition	0.179^{***}	0.179^{***}
	p = 0.000	p = 0.000
Survey_checkout_diff	-0.003^{**}	
	p = 0.023	
Observations	78,440	78,440
R^2	0.671	0.671
Adjusted R^2	0.671	0.671
Note:	*p<0.1; **p<0.05; ***p<0.01	

Results

Variable importance is indicated by the value of each coefficient's test statistic. So we sort the coefficient estimates on the test statistic.

For each of the survey questions below, we can say that a one-unit increase in the variable is associated with a β_k increase in response to the question about recommending Hertz. For example, a one-unit increase in reponse to the question on `Value_for_the_Money` is associated with a 0.29 increase in the question to about recommending Hertz.

```
tidy(lm2) %>%
  as_tibble() %>%
  filter(term != "(Intercept)") %>%
  arrange(desc(statistic)) %>%
  select(term, statistic, estimate) %>%
  pander(justify = c('left', 'right', 'right'))
```

term	statistic	estimate
Value_for_the_Money	100.9	0.2943
Veh_Equip_Condition	77.86	0.1788

term	statistic	estimate
Speed_of_Service	72.9	0.1977
Staff_Courtesy	65.5	0.2395
Trans_Billing_as_Expected	61.82	0.161
Total_charge_USD	8.03	0.0001355

B

To test if the relationships change by Rental Location Type, Rental Purpose, and Booking Channel, we individual add each to the base model. The results are shown in ??:

```
lm3 <- update(lm2, . ~ . + rent_loc_type)
lm4 <- update(lm2, . ~ . + Purpose_of_Rental)
lm5 <- update(lm2, . ~ . + as.factor(booking_channel_dummy))
regression_results(lm2, lm3, lm4, lm5,
                   title = "Rental Location Type, Rental Purpose, and Booking Channel")
```

1. **Rental Location Type**: Yes - all survey question responses held constant, picking up the rental car at the airport increases the response about recommending Hertz by 0.044 points. In reality, this would not translate into a full point.
2. **Rental Purpose**: No - all survey question responses held constant, the purpose of the rental has not impact.
3. **Booking Channel**: Yes - all survey question responses held constant, booking through hertz.com increases the response about recommending Hertz by 0.104 points. In reality, this would not translate into a full point.

C

Useful features to segment customers on need to meet the following criteria:

1. intrinsic to the customer or the customer's experience with Hertz
2. contain variation (i.e `rent_corp_lic` is split 95% and 5%)
3. not contain too many levels (i.e. segmenting on all US/CN states would not be helpful)

We will explore a few in this dataset:

Customer Tier

```
hertz_data %>%
  group_by(cust_tier_code) %>%
  summarise(n = n(), mean = mean(Recom_mend_Hertz, na.rm = TRUE)) %>%
  arrange(desc(n)) %>%
  pander()
```

cust_tier_code	n	mean
RG	46,647	7.572
FG	15,763	7.679
N1	11,390	7.553
PC	5,958	7.934

cust_tier_code	n	mean
PL	703	8.28
NA	251	7.394
VP	8	8.125
PS	3	8.667

```
lm6 <- update(lm2, . ~ . + cust_tier_code)
anova(lm6, lm(lm2$call, data = lm6$model)) %>%
  pander(missing = "")
```

Table 8: Analysis of Variance Table

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
78,189	114,553				
78,195	114,683	-6	-129.5	14.73	6.795e-17

Country

```
hertz_data %>%
  group_by(addr_country) %>%
  summarise(n = n(), mean = mean(Recom_mend_Hertz, na.rm = TRUE)) %>%
  arrange(desc(n)) %>%
  head(10) %>%
  pander()
```

addr_country	n	mean
US	77,251	7.625
CN	2,205	7.565
BR	215	7.851
ME	199	7.809
PR	113	7.522
UK	61	6.787
AR	56	7.661
VN	52	7.654
CO	47	7.596
CR	39	8.231

```
lm7 <- update(lm2, . ~ . + addr_country)
anova(lm7, lm(lm2$call, data = lm7$model)) %>%
  pander(missing = "")
```

Table 10: Analysis of Variance Table

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
78,361	114,985				
78,433	115,158	-72	-172.3	1.631	0.0005823

Table 6: Rental Location Type, Rental Purpose, and Booking Channel

	Dependent variable:			
	Recom_mend_Hertz			
	(1)	(2)	(3)	(4)
Constant	-0.587^{***}	-0.584^{***}	-0.608^{***}	-0.635^{***}
	p = 0.000	p = 0.000	p = 0.000	p = 0.000
Staff_Courtesy	0.239^{***}	0.241^{***}	0.239^{***}	0.239^{***}
	p = 0.000	p = 0.000	p = 0.000	p = 0.000
Speed_of_Service	0.198^{***}	0.198^{***}	0.198^{***}	0.198^{***}
	p = 0.000	p = 0.000	p = 0.000	p = 0.000
Trans_Billing_as_Expected	0.161^{***}	0.161^{***}	0.162^{***}	0.161^{***}
	p = 0.000	p = 0.000	p = 0.000	p = 0.000
Value_for_the_Money	0.294^{***}	0.294^{***}	0.293^{***}	0.294^{***}
	p = 0.000	p = 0.000	p = 0.000	p = 0.000
Total_charge_USD	0.0001^{***}	0.0001^{***}	0.0001^{***}	0.0001^{***}
	p = 0.000	p = 0.000	p = 0.000	p = 0.000
Veh_Equip_Condition	0.179^{***}	0.179^{***}	0.179^{***}	0.179^{***}
	p = 0.000	p = 0.000	p = 0.000	p = 0.000
rent_loc_typeOFF AP		-0.023^{***}		
		p = 0.009		
Purpose_of_RentalIns. Rep. or Loaner			0.114^{**}	
			p = 0.042	
Purpose_of_RentalLeis. / Pers.			0.059^{***}	
			p = 0.000	
as.factor(booking_channel_dummy)1				0.104^{***}
				p = 0.000
Observations	78,440	76,157	78,440	78,440
R^2	0.671	0.671	0.671	0.671
Adjusted R^2	0.671	0.671	0.671	0.671
Note:				*p<0.1; **p<0.05; ***p<0.01

Rental Day

```
hertz_data %>%
  select(Recom_mend_Hertz,  rent_day) %>%
  group_by(rent_day) %>%
  summarise(n = n(), mean = mean(Recom_mend_Hertz, na.rm = TRUE)) %>%
  arrange(rent_day) %>%
  pander()
```

rent_day	n	mean
1	7,021	7.486
2	15,967	7.572
3	12,630	7.58
4	11,234	7.587
5	11,683	7.714
6	13,755	7.735
7	8,433	7.636

```
lm8 <- update(lm2, . ~ . + rent_day)
anova(lm8, lm(lm2$call, data = lm8$model)) %>%
  pander(missing = "")
```

Table 12: Analysis of Variance Table

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
78,432	115,155				
78,433	115,158	-1	-2.079	1.416	0.2341

Difference in Car Reserved and Car Given

```
hertz_data %>%
  mutate(is_same = if_else(xgra_veh_class == xgra_vclass_reserv, "Same", "Different")) %>%
  group_by(is_same) %>%
  summarise(n = n(), mean = mean(Recom_mend_Hertz, na.rm = TRUE)) %>%
  pander()
```

is_same	n	mean
Different	61,135	7.642
Same	19,588	7.563

```
lm9 <- lm(update(lm2$call$formula, ~. + is_same), data = hertz_data %>%
             mutate(is_same = if_else(xgra_veh_class == xgra_vclass_reserv, "Same", "Different")))
anova(lm9, lm2) %>%
  pander(caption = "ANOVA for model with", missing = "")
```

Table 14: ANOVA for model with

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
78,432	115,156				
78,433	115,158	-1	-1.242	0.8459	0.3577

D

1. The segmentation exercise only involved customers who had completed a survey. The respondents likely have a more positive view of Hertz than the typical customer. Rather than segment based on someone's response "how likely are you to recommend", Hertz could track the referrals people make to others as this represents behavior rather than perceived intent.
2. We are segmenting customers based on their responses to other survey questions. This bias does not really help make this segmentation actionable as it's merely descriptive. This analysis only helps tell us that "value for money" is driver in someone's propensity to recommend, but it does not enable future targeting. Segmentation based only on customer characteristics (i.e. age, location, income, vehicle use-case) is much more actionable.
3. We treated an ordinal discrete variables (survey responses on a 1-9 scale) as a continuous variable. We could have performed the analysis using parametric tests that are more suitable to likert scale data.