**LINEAR REGRESSION**

> summary(model\_bl)

Call:

lm(formula = log\_sales ~ . - Item\_Identifier - Outlet\_Identifier -

Item\_Type - Item\_Fat\_Content, data = train)

Residuals:

Min 1Q Median 3Q Max

-0.92052 -0.11864 0.02406 0.15816 0.58918

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.5703169 0.1110211 5.137 2.85e-07 \*\*\*

Item\_Weight -0.0002450 0.0005765 -0.425 0.67091

Item\_Visibility 0.0030785 0.0492743 0.062 0.95018

log\_MRP 1.0226641 0.0106595 95.940 < 2e-16 \*\*\*

Establishment\_Age -0.0057564 0.0020739 -2.776 0.00552 \*\*

Outlet\_SizeMedium -0.1283900 0.0546213 -2.351 0.01877 \*

Outlet\_SizeSmall -0.1342582 0.0507016 -2.648 0.00811 \*\*

Outlet\_Location\_TypeTier 2 -0.0332119 0.0174235 -1.906 0.05666 .

Outlet\_Location\_TypeTier 3 -0.0854163 0.0302276 -2.826 0.00473 \*\*

Outlet\_TypeSupermarket Type1 0.7687520 0.0278457 27.608 < 2e-16 \*\*\*

Outlet\_TypeSupermarket Type2 0.7070607 0.0246346 28.702 < 2e-16 \*\*\*

Outlet\_TypeSupermarket Type3 1.1528095 0.0350047 32.933 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2248 on 8511 degrees of freedom

Multiple R-squared: 0.7415, Adjusted R-squared: 0.7411

F-statistic: 2219 on 11 and 8511 DF, p-value: < 2.2e-16

> summary(model1)

Call:

lm(formula = log\_sales ~ ., data = train.data)

Residuals:

Min 1Q Median 3Q Max

-0.90985 -0.12056 0.02357 0.15871 0.57535

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.150e-01 1.254e-01 4.905 9.58e-07 \*\*\*

Item\_Weight -5.401e-05 6.472e-04 -0.083 0.93350

Item\_Fat\_ContentRegular 1.133e-02 6.304e-03 1.798 0.07230 .

Item\_Visibility 1.453e-02 5.518e-02 0.263 0.79235

Item\_TypeBreads 2.098e-02 1.862e-02 1.127 0.25988

Item\_TypeBreakfast -4.328e-02 2.556e-02 -1.694 0.09040 .

Item\_TypeCanned 1.162e-02 1.412e-02 0.823 0.41064

Item\_TypeDairy -1.629e-02 1.399e-02 -1.164 0.24454

Item\_TypeFrozen Foods -1.586e-02 1.331e-02 -1.192 0.23320

Item\_TypeFruits and Vegetables -7.856e-03 1.233e-02 -0.637 0.52402

Item\_TypeHard Drinks -2.759e-03 2.006e-02 -0.138 0.89061

Item\_TypeHealth and Hygiene -1.597e-03 1.531e-02 -0.104 0.91693

Item\_TypeHousehold -1.209e-02 1.340e-02 -0.902 0.36713

Item\_TypeMeat 5.074e-03 1.567e-02 0.324 0.74620

Item\_TypeOthers -4.763e-03 2.206e-02 -0.216 0.82910

Item\_TypeSeafood -9.643e-03 3.309e-02 -0.291 0.77077

Item\_TypeSnack Foods -1.055e-02 1.242e-02 -0.849 0.39597

Item\_TypeSoft Drinks -6.169e-04 1.584e-02 -0.039 0.96894

Item\_TypeStarchy Foods -8.900e-03 2.299e-02 -0.387 0.69863

log\_MRP 1.024e+00 1.211e-02 84.538 < 2e-16 \*\*\*

Establishment\_Age -6.852e-03 2.338e-03 -2.930 0.00340 \*\*

Outlet\_SizeMedium -1.533e-01 6.153e-02 -2.491 0.01275 \*

Outlet\_SizeSmall -1.547e-01 5.709e-02 -2.710 0.00674 \*\*

Outlet\_Location\_TypeTier 2 -4.404e-02 1.957e-02 -2.250 0.02447 \*

Outlet\_Location\_TypeTier 3 -9.029e-02 3.402e-02 -2.654 0.00798 \*\*

Outlet\_TypeSupermarket Type1 7.571e-01 3.134e-02 24.157 < 2e-16 \*\*\*

Outlet\_TypeSupermarket Type2 6.947e-01 2.763e-02 25.142 < 2e-16 \*\*\*

Outlet\_TypeSupermarket Type3 1.167e+00 3.940e-02 29.609 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2251 on 6790 degrees of freedom

Multiple R-squared: 0.739, Adjusted R-squared: 0.738

F-statistic: 712.2 on 27 and 6790 DF, p-value: < 2.2e-16

> vif(model1)

GVIF Df GVIF^(1/(2\*Df))

Item\_Weight 1.015914 1 1.007925

Item\_Fat\_Content 1.217129 1 1.103236

Item\_Visibility 1.100551 1 1.049071

Item\_Type 1.271411 15 1.008036

log\_MRP 1.013341 1 1.006648

Establishment\_Age 51.303922 1 7.162676

Outlet\_Size 175.686630 2 3.640698

Outlet\_Location\_Type 112.360915 2 3.255771

Outlet\_Type 484.813241 3 2.802823

> summary(model2)

Call:

lm(formula = log\_sales ~ ., data = train.data)

Residuals:

Min 1Q Median 3Q Max

-0.90563 -0.11925 0.02295 0.15915 0.58671

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.5994861 0.1249848 4.796 1.65e-06 \*\*\*

Item\_Weight -0.0002003 0.0006428 -0.312 0.75538

Item\_Fat\_ContentRegular 0.0119136 0.0057196 2.083 0.03730 \*

Item\_Visibility 0.0088607 0.0549581 0.161 0.87192

log\_MRP 1.0229936 0.0120360 84.994 < 2e-16 \*\*\*

Establishment\_Age -0.0066060 0.0023358 -2.828 0.00470 \*\*

Outlet\_SizeMedium -0.1468889 0.0614668 -2.390 0.01689 \*

Outlet\_SizeSmall -0.1485441 0.0570369 -2.604 0.00922 \*\*

Outlet\_Location\_TypeTier 2 -0.0420837 0.0195510 -2.153 0.03139 \*

Outlet\_Location\_TypeTier 3 -0.0868953 0.0339893 -2.557 0.01059 \*

Outlet\_TypeSupermarket Type1 0.7603070 0.0313077 24.285 < 2e-16 \*\*\*

Outlet\_TypeSupermarket Type2 0.6971899 0.0276065 25.255 < 2e-16 \*\*\*

Outlet\_TypeSupermarket Type3 1.1635130 0.0393678 29.555 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2251 on 6805 degrees of freedom

Multiple R-squared: 0.7385, Adjusted R-squared: 0.7381

F-statistic: 1602 on 12 and 6805 DF, p-value: < 2.2e-16

> model2 <- lm(log\_sales ~ .-Item\_Identifier-

+ Outlet\_Identifier-Establishment\_Age,data = train\_new)

> summary(model2)

Call:

lm(formula = log\_sales ~ . - Item\_Identifier - Outlet\_Identifier -

Establishment\_Age, data = train\_new)

Residuals:

Min 1Q Median 3Q Max

-0.91636 -0.12043 0.02282 0.15800 0.57947

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.2683132 0.0396749 6.763 1.47e-11 \*\*\*

Item\_Weight -0.0001891 0.0006533 -0.289 0.7722

Item\_Fat\_ContentRegular 0.0032992 0.0062861 0.525 0.5997

Item\_Visibility 0.0267646 0.0555325 0.482 0.6298

Item\_TypeBreads 0.0041136 0.0185923 0.221 0.8249

Item\_TypeBreakfast -0.0289665 0.0259391 -1.117 0.2642

Item\_TypeCanned -0.0026236 0.0141569 -0.185 0.8530

Item\_TypeDairy -0.0237088 0.0139849 -1.695 0.0901 .

Item\_TypeFrozen Foods -0.0180035 0.0131733 -1.367 0.1718

Item\_TypeFruits and Vegetables -0.0031072 0.0123901 -0.251 0.8020

Item\_TypeHard Drinks -0.0056210 0.0208000 -0.270 0.7870

Item\_TypeHealth and Hygiene 0.0018822 0.0153133 0.123 0.9022

Item\_TypeHousehold -0.0164947 0.0135086 -1.221 0.2221

Item\_TypeMeat 0.0005957 0.0157777 0.038 0.9699

Item\_TypeOthers -0.0165054 0.0224073 -0.737 0.4614

Item\_TypeSeafood -0.0173404 0.0328298 -0.528 0.5974

Item\_TypeSnack Foods -0.0116934 0.0123912 -0.944 0.3454

Item\_TypeSoft Drinks -0.0141423 0.0155992 -0.907 0.3647

Item\_TypeStarchy Foods -0.0123865 0.0228543 -0.542 0.5879

log\_MRP 1.0280162 0.0121681 84.484 < 2e-16 \*\*\*

Outlet\_SizeMedium 0.0183293 0.0192803 0.951 0.3418

Outlet\_SizeSmall 0.0020868 0.0193094 0.108 0.9139

Outlet\_Location\_TypeTier 2 0.0116866 0.0095786 1.220 0.2225

Outlet\_Location\_TypeTier 3 -0.0096998 0.0153927 -0.630 0.5286

Outlet\_TypeSupermarket Type1 0.8349026 0.0140797 59.298 < 2e-16 \*\*\*

Outlet\_TypeSupermarket Type2 0.7549526 0.0179348 42.094 < 2e-16 \*\*\*

Outlet\_TypeSupermarket Type3 1.0663380 0.0179411 59.435 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2253 on 6791 degrees of freedom

Multiple R-squared: 0.7384, Adjusted R-squared: 0.7374

F-statistic: 737.3 on 26 and 6791 DF, p-value: < 2.2e-16

> vif(model2)

GVIF Df GVIF^(1/(2\*Df))

Item\_Weight 1.017285 1 1.008605

Item\_Fat\_Content 1.208197 1 1.099180

Item\_Visibility 1.105892 1 1.051614

Item\_Type 1.258775 15 1.007701

log\_MRP 1.012245 1 1.006104

Outlet\_Size 16.963440 2 2.029451

Outlet\_Location\_Type 14.074290 2 1.936897

Outlet\_Type 17.118901 3 1.605385

**Comparison of Linear Regression & Random Forest Models**

|  |  |
| --- | --- |
|  | **RMSE** |
| **Linear Model** | 1101.07 |
| **Random Forest Model 1** | 1119.03 |
| **Random Forest Model 2** | 1125.88 |

# All the calculations are in attached excel files.