Our Planet

Milestone 4

2015 to 2020 Strategy

Colgate's 2015 to 2020 Sustainability Strategy maintains our emphasis on People, Performance and Planet with focused, measurable goals that align with the Company's business objectives.

- Helping Colgate People and Their Families Live Better
- Contributing to the Communities Where We Live and Work
- Delighting Consumers and Sustaining Our World With Our Brands
- Making Every Drop of Water Count
- Reducing Our Impact on Climate and the Environment

WE WILL:

- Responsibly source forest commodities to reach zero net deforestation
- Promote use of renewable energy and reduce absolute greenhouse gas emissions from manufacturing by 25% compared to 2002
- Reduce our manufacturing energy intensity by one-third compared to 2002
- Halve our manufacturing waste sent to landfill per ton of product compared to 2010, working toward our goal of "Zero Waste"
- Partner with key suppliers, customers and consumers to reduce energy, greenhouse gas emissions, and waste

Calvin Sparrow • 08.18.2020

[1] 15% risk reduction will be measured considering a 2013 baseline, using the Global Health Risk Assessment tool, available to countries with 100 or more employees.

[2] The performance results will be based on representative new products and product updates evaluated against comparable Colgate products, considering a 2015 baseline, across seven impact areas to characterize likely improvement in the sustainability profile, based on review of quantitative and qualitative data.

[3] Packages meeting all three criteria are considered recyclable: 1) the package is made of a material that is widely accepted for recycling, 2) the package can be separated into material(s) that can be recycled, and 3) the package material can be reprocessed into a preferred valuable feedstock.

Our Planet

Contents

Review of Questions to Answer / Hypothesis / Approach

- **Recent Progress**
- **Discuss Technical Challenges**

  Detail: ERD

- **Initial Findings**
- **Deeper Analysis**
- **Hypothesis Results**

Our Planet

Questions To Answer -

Why Colgate-Palmolive?

Luck of the alphabet. This project was the offshoot of another project that labeled Indonesia as an area ready for environmental change. So I looked into companies operating in Indonesia.

What is deforestation?

Deforestation refers to the cutting, clearing, and removal of rainforest or related ecosystems into less bio-diverse ecosystems such as pasture, cropland, or plantations.

What has been Colgate's efforts?

In 2013 a self imposed NDPE policy^

Throughout 2016 -2018 there were several attempts by Several NGOs, Greenpeace and RSPO* to get the big Consumer brands and palm oil traders on the same side.

Finally in 2018 Greenpeace challenged leading consumer brands to demonstrate their progress towards responsible sourcing by revealing the mills that produced their palm oil and the names of the producer groups that controlled those mills

Initial Hypothesis -

Too many different sources of data all indicate the same thing. That the level of deforestation in Indonesia has not dropped despite claims by multinational firms to the contrary.

Looking at the graph on the right, you can see that even with a dramatic drop off from 2016, the level of tree cover loss still resulted in a doubling in deforestation from 2001 to 2018.

Our Planet

Section 3: Data Analysis Approach

- Multiple ways were needed to explain what was seen when looking at the numbers, and I had to organize the numbers so the data was easy to get to and not the way it was presented to me.

- There are 33 provinces in the country of Indonesia. The data I have covers the years 2001 - 2018, the most current release of data. There are 3 columns of data; whrc_aboveground_biomass_loss_Mg, whrc_aboveground_co2_emissions, umd_treecover_loss_ha. To decipher this, whrc - Woods Hole Research Center, Mg - Millions of gallons, umd - University of Maryland, ha - hectares

- To look for a trend a started doing a single regression of the whrc_aboveground_biomass_loss_Mg against Years for each individual province

- I still needed to see something graphic so I went to app.rawgraphs.io and used a Circle Packing Graph to draw out the levels of differences in aboveground_biomass_loss over the 18 years and see if any trends caught my eye.

- Finally I ran a Multiple regression of all three columns of data against Years

Our Planet

Technical Challenges -

- Surprisingly not SQL friendly. It’s not that there was a lot of missing data or malformed data, it was the way it was structured/formatted and the volume of different sources.

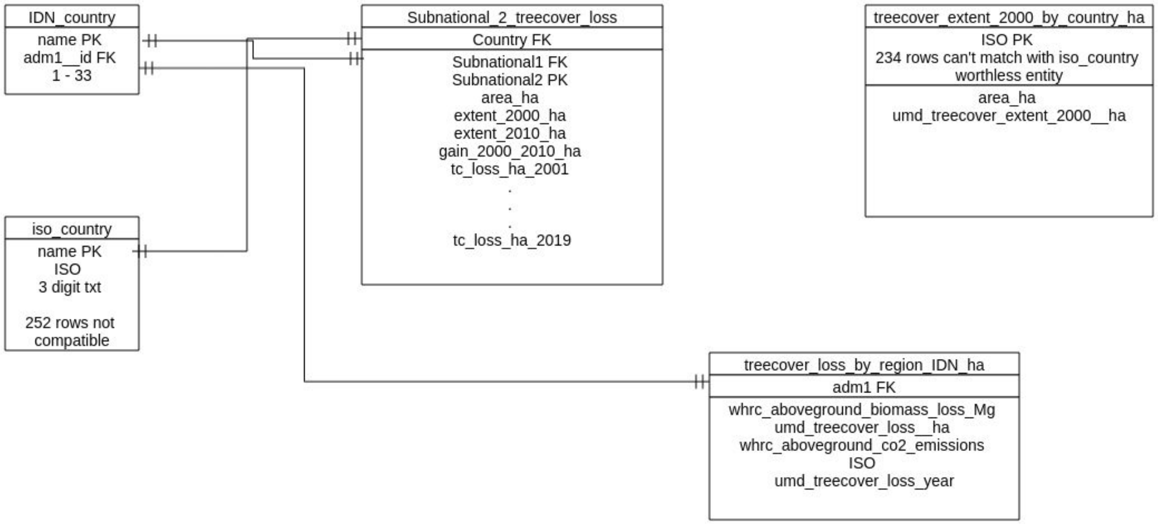
- It took time to come up with a working ERD diagram, and then I ended up using only half.

- It was hard to remain impartial and non-judgmental especially after wasting a couple of days falling down a “rabbit hole”. But it was information worth learning.

- The list of country codes does not match the ISO country listing so you can’t have any Key relationship

- I had to create a reusable block of SQL code that I could run on my one table that had good data treecover_loss_by_region_IDN_ha, to come up with a viable set of 33 pandas dataframes to use throughout this project

Technical Challenges - ERD



Our Planet

Technical Challenges

treecover_loss_by_region_IDN_ha
adm1 FK
whrc_aboveground_biomass_loss_Mg
umd_treecover_loss__ha
whrc_aboveground_co2_emissions
ISO
umd_treecover_loss_year

+

```sql
SELECT round(cast(("umd_tree_cover_loss__ha") AS
NUMERIC),2) AS und_tree_cover_loss_ha
FROM "treecover_loss_by_region_IDN_ha"
WHERE adm1 = 33 and umd_tree_cover_loss__year in
(2001,2002,2003,2004,2005,2006,2007,2008,2009,201
0,2011,2012,1013,2014,2015,2016,2017,2018,
2019);
```

=

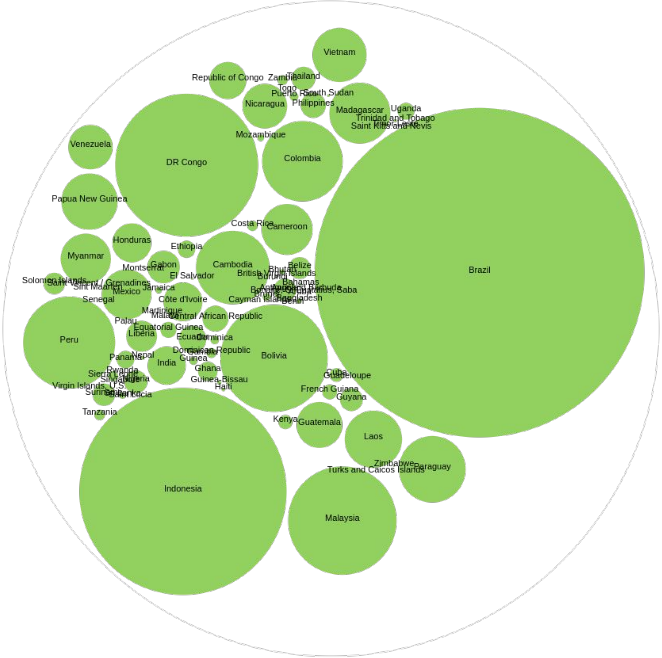
Yogyakarta	whrc_aboveground_biomass_loss_Mg	whrc_aboveground_co2_emissions	umd_treecover_loss_ha
2001	18133.51	33244.77	82.86
2002	6763.83	12400.36	32.31
2003	1822.81	3341.82	8.54
2004	13455.82	24669.01	70.42
2005	13075.74	23972.19	61.73
2006	3893.92	7138.85	19.59
2007	5414.03	9925.72	27.74

Our Planet

Initial Findings - app.rawgraphs.io

Gorontalo on the left Showing aboveground_biomass_loss_Mg for each year. I have an image like this for each of the 33 provinces.

On the right is a global graph of the level of deforestation in 2018. You will notice Indonesia is not far from Brazil if you add in its neighbors



Our Planet

Initial Findings - Regression Analysis

Once I got my individual Datasets I could visualize The trend in the data just by looking at the numbers and see my hypothesis was right, I just had to prove it. Regression Analysis was the best tool to prove a linear trend.

I chose to analyze the whrc-aboveground_biomass_loss_Mg against Year The results were not as encouraging as I would have hoped. Out of 33 runs, 7 had a positive score between .22 and .55

Dataset Cards

Bengkulu	whrc_aboveground_biomass_loss_Mg	whrc_aboveground_co2_emissions	umd_treecover_loss_ha
2001	3884943.39	7122396.22	14198.56
2002	2787589.77	5110581.25	10421.96
2003	2266452.46	4155162.84	8103.87
2004	6999386.89	12832209.3	25984.82
2005	5947585.83	10903907.35	23588.43
2006	4566146.09	8371267.84	17580.27
2007	5738301.4	10520219.23	22353.24
2008	5021281.95	9205683.57	18755.75
2009	4074570.01	7470045.01	16012.66
2010	6545479.28	12000045.34	24562.81
2011	7474045.46	13702416.68	27761.15
2012	3737937.63	6852885.65	14152.78
2013	5161838.99	9463371.47	20319.73
2014	5285359.12	9689825.05	20197.3
2015	6231618.71	11424634.3	24377.83
2016	6959689.16	12759430.13	27232.96
2017	6199934.6	11366546.77	24624.47
2018	4383925.45	8037196.66	18200.56

			treecover_loss_ha
			1054.38
			1761.43
			293.14
			1404.16
			1938.21
			1061.44
2007	202015.6	370361.94	929.75
2008	616341.95	1129960.24	2754.78
2009	360737.78	661352.6	1632.16
2010	480297.15	880544.78	2327.97
2011	737131.36	1351407.5	3286.61
2012	368595.05	675757.58	1830.15
2013	464203.43	851039.62	2274.87
2014	515669.01	945393.18	2663.78
2015	390136.02	715249.37	1970.35
2016	680050.98	1246760.13	3495.04
2017	667984.4	1224638.07	3400.14
2018	876177.4	1606325.23	4714.45

			ecover_loss_ha
			13909.72
			11247.04
			7123.08
			21976.07
			20241.32
			23237.46
			22677.86
			63895.32
			23187.37
			32399.71
			39450.87
2012	3206112.27	5877872.49	17575.17
2013	7205800.5	13210634.25	41245.15
2014	7071565.28	12964536.34	41280.97
2015	9337860.84	17119411.54	58202.69
2016	4416974.12	8097785.89	25557.97
2017	3661925.09	6713529.33	22425.66
2018	4172583.12	7649735.72	25528.89

OBJ

Our Planet

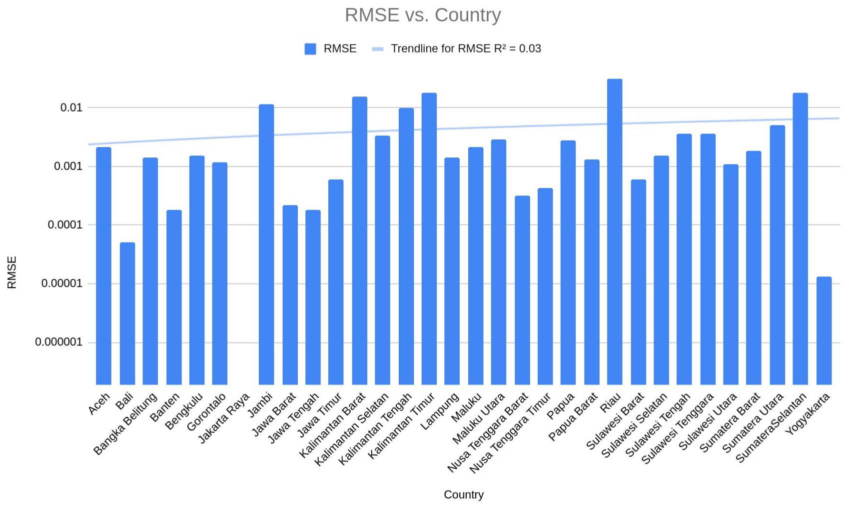
Initial Findings - Regression Analysis 2

Root Mean Square Error - Is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far away from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit.

$$RMSE = \sqrt{(f - o)^2}$$

- f = forecasts (expected values or unknown results),
- o = observed values (known results).

Range of RMSE is 13,213.67 - 31,310,374.39



Our Planet

Deeper Analysis

- The results are showing me the same things. Although the individual numbers are still scattered which represents random logging and deforestation activities, the overall trend since peaking in the middle of the decade is slowly moving down.

- I used the Circle Packing chart from app.RAWGraphs.io Nested circles allow me to represent hierarchies and compare values.
- This visualization is particularly effective to show the proportion between elements through their areas and their position inside a hierarchical structure. Based on http://bl.ocks.org/mbostock/4063530

- Since I have three variables whrc_aboveground_biomass_loss_Mg, whrc_aboveground_co2_emissions, and umd_treecover_loss_ha, to measure against one constant, the number of years 2001 - 2018 I will do a multiple regression on all the datasets also keeping track of the model score and RMSE.

- All these data files are available on GitHub in sets of 3. 1 .png file 2 .html files for each province. http://github.com/csparrow99/practice

Our Planet

Deeper Analysis - Multiple Regression

```
In [219]: dataset.describe()
```

Out[219]:

	Year	whrc_aboveground_biomass_loss_Mg	whrc_aboveground_co2_emissions	umd_treecover_loss_ha
count	18.000000	18.000000	18.000000	18.000000
mean	2009.500000	13611.717222	24954.813333	67.102778
std	5.338539	11235.853606	20599.063244	46.377157
min	2001.000000	1822.810000	3341.820000	8.540000
25%	2005.250000	5751.480000	10544.380000	28.882500
50%	2009.500000	11227.795000	20584.285000	59.745000
75%	2013.750000	16964.087500	31100.830000	79.750000
max	2018.000000	42945.970000	78734.270000	173.010000

```
In [220]: X= dataset[['whrc_aboveground_biomass_loss_Mg', 'whrc_aboveground_co2_emissions', 'umd_treecover_loss_ha']]
          y = dataset['Year']

In [221]: from sklearn.model_selection import train_test_split
          X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)

In [222]: from sklearn.linear_model import LinearRegression
          regressor = LinearRegression()
          regressor.fit(X_train, y_train)

Out[222]: LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None, normalize=False)

In [223]: coeff_df = pd.DataFrame(regressor.coef_, X.columns, columns=['Coefficient'])
          coeff_df
```

3	2004	13455.82	24669.01	70.42
4	2005	13075.74	23972.19	61.73


```
In [219]: dataset.describe()
```

Out[219]:

	Year	whrc_aboveground_biomass_loss_Mg	whrc_aboveground_co2_emissions	umd_treecover_loss_ha
count	18.000000	18.000000	18.000000	18.000000
mean	2009.500000	13611.717222	24954.813333	67.102778
std	5.338539	11235.853606	20599.063244	46.377157
min	2001.000000	1822.810000	3341.820000	8.540000
25%	2005.250000	5751.480000	10544.380000	28.882500
50%	2009.500000	11227.795000	20584.285000	59.745000
75%	2013.750000	16964.087500	31100.830000	79.750000
max	2018.000000	42945.970000	78734.270000	173.010000

```
In [220]: X= dataset[['whrc_aboveground_biomass_loss_Mg', 'whrc_aboveground_co2_emissions', 'umd_treecover_loss_ha']]
          y = dataset['Year']

In [221]: from sklearn.model_selection import train_test_split
          X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)

In [222]: from sklearn.linear_model import LinearRegression
          regressor = LinearRegression()
          regressor.fit(X_train, y_train)

Out[222]: LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None, normalize=False)

In [223]: coeff_df = pd.DataFrame(regressor.coef_, X.columns, columns=['Coefficient'])
          coeff_df
```


Out[223]:

	Coefficient
whrc_aboveground_biomass_loss_Mg	1158.705467
whrc_aboveground_co2_emissions	-632.021848
umd_treecover_loss_ha	0.292430

```
In [224]: # to make predictions on the test data
          y_pred = regressor.predict(X_test)

In [225]: print(regressor.score(X_test,y_test))

0.08074523996041438

In [226]: df = pd.DataFrame({'Actual': y_test, 'Predicted': y_pred})
          df
```

Out[226]:

	Actual	Predicted
1	2002	2003.667188
6	2007	2008.238195
8	2009	2003.067807
10	2011	2012.277927

```
In [227]: from sklearn import metrics
          print('Mean Absolute Error:', metrics.mean_absolute_error(y_test, y_pred))
          print('Mean Squared Error:', metrics.mean_squared_error(y_test, y_pred))
          print('Root Mean Squared Error:', np.sqrt(metrics.mean_squared_error(y_test, y_pred)))

Mean Absolute Error: 2.5288755663328857
Mean Squared Error: 10.284162627942864
Root Mean Squared Error: 3.2068929866683833
```

Our Planet

Deeper Analysis - Multiple Regression 2

- Following the flow from the previous page, after the dataset is described, train_test_split is imported from model selection and then LinearRegression
- regressor=LinearRegression() regressor.fit(X_train,y_train) X_train, y_train came from train_test_split
- Now the model has been trained, and has given us these Coefficients Out[358]:
- Here is our Model Score 0.6546 or 65.46% accuracy in predicting the combined direction of the 3 variables at any given yr.
- In [359]: is used to predict the test data in In [361]
- Note the low RMSE relative to the high Model score and turn the page!

Out[358]:

	Coefficient
whrc_aboveground_biomass_loss_Mg	636.643395
whrc_aboveground_co2_emissions	-347.260036
umd_treecover_loss_ha	0.001286

In [359]:
```
# to make predictions on the test data
y_pred = regressor.predict(X_test)
```

In [360]:
```
print(regressor.score(X_test,y_test))
```

```
0.654617193603948
```

In [361]:
```
df = pd.DataFrame({'Actual': y_test, 'Predicted': y_pred})
df
```

Out[361]:

	Actual	Predicted
1	2002	2004.291125
6	2007	2004.115619
8	2009	2010.192189
10	2011	2010.317608

In [362]:
```
from sklearn import metrics
print('Mean Absolute Error:', metrics.mean_absolute_error(y_test, y_pred))
print('Mean Squared Error:', metrics.mean_squared_error(y_test, y_pred))
print('Root Mean Squared Error:', np.sqrt(metrics.mean_squared_error(y_test, y_pred)))
```

```
Mean Absolute Error: 1.7625217428137603
Mean Squared Error: 3.863970146555832
Root Mean Squared Error: 1.9656983864662025
```

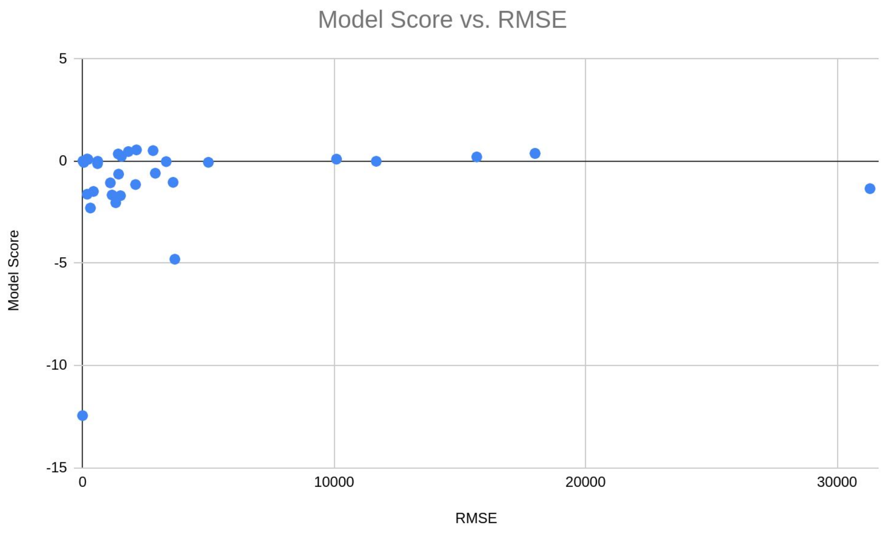
Deeper Analysis - New Metric?

- The single linear regression analysis didn’t reveal anything really useful, eyeballing the statics was just as productive.
- The multiple regression analysis told a different story. There was an obvious correlation between something not obvious.
- When I added one line of code: `print(regressor.score(X_test,y_test))` and ran all 33 regressions again to get a model score I noticed a definite correlation between -
- RMSE and Model Score
- With the single linear regression the Model Score clustered around 0 - (-2) no matter what the RMSE
- However with the multiple regression the Model Score dropped on a very smooth line with a rise in the RMSE
- The 2 graphs in the following slide show the difference.

Our Planet

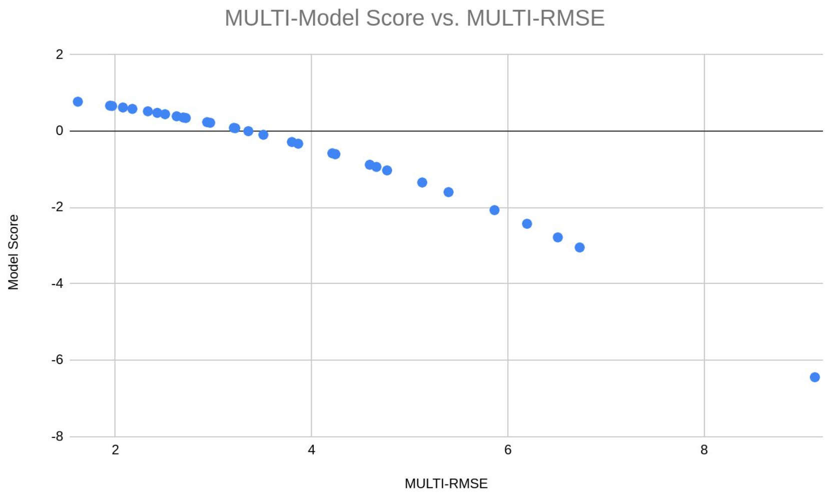
Graphs

Model Score vs. RMSE Single Regression



MULTI-Model Score vs MULTI-RMSE

Multiple Regression



Final Findings (Results of Hypothesis)

- Initial Hypothesis: Too many different sources of data all indicate the same thing. That the level of deforestation in Indonesia has not dropped despite claims by multinational firms to the contrary.
- Where we stand today: The levels of deforestation have dropped from their peak of 2014,2015, 2016 but not to the levels that meet some of the promises that were made in the middle of the decade.
- This is most evident in the 33 graphs on http://github.com/csparrow/practice from app.rawgraphs.io which are Circle Packing graphs to show the nested hierarchy of whrc_aboveground_biomass_loss_Mg and Years
- The new metric MULTI-RMSE vs MULTI- Model Score can be useful to determine the legitimacy of your multivariable model

Final Findings - (Results of Hypothesis 2)

- In 2018 Greenpeace published a book “Now Or Never To Reform The Palm Oil Industry” in which they outlined demands for major consumer goods companies to meet in Greenpeace’s quest to clean up the supply chain hence eliminating deforestation among other things.
- “At the same time, however, the supply chain information disclosed by brands and traders indicates that little progress has been made towards cleaning up the global palm oil trade. Every company that has opened its supply chain to public scrutiny is sourcing from producers that are known to be clearing rainforests, exploiting their workers and/or embroiled in land conflicts with local communities.”
- What needs to be done:
- Brands and traders need to take responsibility for screening the producers in their supply chains to ensure they are not doing business with groups that are destroying rainforests. They need their own comprehensive monitoring system, based on their suppliers’ mill location data and concession maps, to ensure that the producer groups in their supply chains comply fully with NDPE standards. Crucially, information regarding producer groups’ landholdings and operations should be placed in the public domain to enable any claims to be independently verified.
- According to Colgate’s website on Sustainability they have started initiatives addressing all these issues, there is a .pdf you can download with all the mills current as of 2017

Our Planet

Yes, This page looks off because something is off

% Gain co2 Emissions By Province 2002 - 2018	
Aceh	139.31
Bali	76.19
Bangka Belitung	89.68
Banten	112.97
Bengkulu	57.26
Gorontalo	16.76
Jakarta Raya	1
Jambi	38.73
Jawa Barat	29.12
Jawa Tengah	13.81
Jawa Timur	-4.41
Kalimantan Barat	29.87
Kalimantan Selatan	62.76
Kalimantan Tengah	1.6
Kalimantan Timur	102.25
Lampung	7.88
Maluku	9.04
Maluku Utara	-43.62
Nusa Tenggara Barat	0
Nusa Tenggara Timur	30.84
Papua	85.31
Papua Barat	16.71
Riau	-46.46
Sulawesi Barat	11.23
Sulawesi Selatan	40.33
Sulawesi Tengah	15.88
Sulawesi Tenggara	106.11
Sulawesi Utara	-9.38
Sumatera Barat	88.52
Sumatera Utara	15.6
Sumatera Selantan	38.08
Yogyakarta	72.22
Average	+37.66

ainability World and North America Indices, was recognized as a U.S. EPA ENERGY STAR 2018 Partner of the Year for the 8th year in a row, and was

Reducing Our Impact on Climate and the Environment

Colgate continues to reduce its absolute greenhouse gas emissions. So far, we have reduced our absolute greenhouse gas emissions by approximately 30% compared to 2002.[(2)]

Working toward the Company's goal of "Zero Waste," Colgate has reduced the amount of waste per ton of production sent to landfills by nearly 41% since 2010.[(2)]

Colgate continues to make progress on its commitment to mobilize resources to achieve zero net deforestation by 2020 as stated in our policy on No Deforestation.

ːe likely improvement in the sustainability profile, based on review of quantitative and qualitative data.